TOPICS IN HIGH-DIMENSIONAL ASYMPTOTICS OF RIDGE-TYPE ESTIMATORS

Bingxin Zhao

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Hongtu Zhu

Joseph G. Ibrahim

Fei Zou

Yuchao Jiang

Jason L. Stein

ABSTRACT

Bingxin Zhao: Topics in high-dimensional asymptotics of ridge-type estimators
(Under the direction of Hongtu Zhu)

In many big data applications, massive features are collected and the number of signals can be large, resulting in a high-dimensional sparsity free (dense) setting, many features are 'true', but weak signals. Genome-wide association studies (GWAS) epitomize this kind of situation, but similar applications widely arise in neuroimaging, omics data, and social science, among others. The goal of this thesis is to model and investigate dense signal problems in a high-dimensional sparsity free framework. Motivated by GWAS applications, we theoretically investigate several problems that are of great practical interest in scientific contexts, leading to better statistical analysis and out-of-sample prediction for complex traits.

Our main results include a random matrix theory framework for GWAS and the associated high-dimensional asymptotic results for a class of ridge-type estimators. These theoretical results help address three aspects in GWAS applications. First, we study the cross-trait polygenic risk score (PRS) method for genetic correlation estimation. We show that, though intuitive and commonly used in practice, the genetic correlation estimated by cross-trait PRS can be asymptotically biased towards zero. We propose a consistent cross-trait PRS estimator to correct such asymptotic bias. Second, empirical evidence has shown that a substantial and widespread gap exists between GWAS signal strength and prediction performance. We investigate this phenomenon for a class of ridge-type estimators, identify the key factors that determine the gap, and illustrate that such gap is a fundamental analytic challenge for all ridge-type estimators in the presence of many true weak signals. Third, we study the assembled ridge estimators, which can be efficiently generated by combining together the

marginal estimator learned from training data and the feature covariance structure estimated on an independent reference panel. For data with a block-diagonal covariance structure, we reveal that the block-wise assembled estimators not only enjoy superior computational efficiency, but can also have very similar performance to the original estimators. We also propose a novel assembled estimator to improve the prediction accuracy. We illustrate our theoretical results by using the simulated data and real GWAS in the UK Biobank database.

ACKNOWLEDGEMENTS

I am grateful to Hongtu Zhu for his guidance, encouragement, and support during my study at UNC. I thank Fei Zou for introducing me to statistical research at UF and for many valuable discussions afterwards. I am also grateful to Joseph G. Ibrahim for supporting my academic activities. Thanks to Yun Li for her collaboration and help on my job search. I also thank Yuchao Jiang and Jason L. Stein for their advice and serving in my committee.

I thank my fellow students, other faculty and staff at UNC Biostatistics for many enjoyable memories. I am grateful to Laura Y. Zhou for her enormous help on my job talk.

I thank Tengfei Li, Xifeng Wang, Yue Yang, Tianyou Luo, Yue Annie Shan, and Jingwen Zhang for intense collaborations in these years. Special thanks to Ziliang Zhu for many conversations on interesting math problems. I have learnt a lot from these wonderful persons.

For the last four years, it has been my pleasure to be a member of the Biostatistics and Imaging Genomics Analysis Lab (BIG-S2). I would like to take this opportunity to thank all my lab mates in BIG-S2 for priceless friendship. I feel very lucky to meet and know them personally and have enjoyed the lab meetings and activities. Additional thanks are due to the staff at UNC Davis library, where I finished most of the work in this dissertation.

Finally and most importantly, I would like to dedicate this dissertation to Jinjie Lin, Xinli Huan, Zhenmin Zhao, all of my friends and the whole family for their love in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Recent developments of big data technologies enable collection of massive features from a large number of individuals. In many real data applications, the number of signals can be large, resulting in a high-dimensional dense (sparsity free) setting, many features are 'true', but weak signals. Genome-wide association studies (GWAS) epitomize this kind of situation, but similar applications widely arise in neuroimaging, omics data, and social science, among others. New empirical challenges have been arising in these dense signal problems, motivating us to model them in a high-dimensional sparsity free framework and investigate several aspects of asymptotics in scientific contexts.

In this thesis, we first study the cross-trait polygenic risk score (PRS) method, which has gained popularity for assessing genetic correlation of complex traits using summary statistics from biobank-scale GWAS. We theoretically show that the estimated genetic correlation by cross-trait PRS can be asymptotically biased towards zero. We propose a consistent cross-trait PRS estimator to correct such asymptotic bias. We also study the variance of cross-trait PRS and explain why the estimator can still be have reliable p-value even it is heavily biased towards zero. In addition, we investigate whether or not SNP screening by GWAS p-values can lead to improved estimation and show the effect of overlapping samples among GWAS. Our results may help demystify and tackle the puzzling "missing genetic overlap" phenomenon of cross-trait PRS for dissecting the genetic similarity of closely related heritable traits.

Second, we propose a general random matrix theory framework to analyze GWAS out-of-sample prediction. Empirical evidence has shown that a substantial and widespread gap exists between GWAS signal strength and performance of marginal summary statistics. We investigate such gap for a class of ridge-type estimators, including the popular marginal

estimator and the best linear unbiased prediction (BLUP) estimator as two special cases. We illustrate that such gap is a fundamental analytic challenge for all ridge-type estimators in the presence of many true weak signals. Furthermore, we show that the relative out-of-sample performance of these estimators highly depends on $\omega = p/n$, the ratio of dimension p over sample size n. Particularly, it reveals that marginal estimator can easily become near-optimal within this class when ω is large, even though it is an extremely over-regularized one. On the other hand, BLUP estimator can become substantially better than marginal estimator as ω is close to one. Our results have important implications in GWAS and other non-sparse problems.

Finally, we study the assembled ridge estimators, which can be efficiently generated by combining together the marginal estimator learned from training data and the feature covariance structure estimated on an independent reference panel. We investigate the relative performance of assembled estimators and original estimators that are directly trained from training data. Moreover, for block-diagonal covariance structure, we reveal that the assembled estimators not only enjoys superior computational efficiency, but also has very similar performance to the original estimators. Based on our theoretical results, we also propose a novel assembled estimator to improve the prediction accuracy. We illustrate our theoretical results by using the simulated data and real GWAS in the UK Biobank database.

CHAPTER 2: LITERATURE REVIEW

2.1 Genome-wide association studies

Human complex traits often have a polygenic genetic architecture (O'Connor et al., 2019; Wray et al., 2018; Boyle et al., 2017). That is, a large number of genetic variants have small but nonzero contributions to phenotypic variation (Timpson et al., 2018). Genome-wide association studies (GWAS) aim to find suspicious genetic risk variants by examining association between complex traits and millions of variants, typically common (minor allele frequency [MAF] ≥ 0.05) single-nucleotide polymorphisms (SNPs) collected across the genome. After a decade of GWAS discovery, more than 100 millions of individuals have been genotyped (Martin et al., 2019) and thousands of unique traits have been studied (Visscher et al., 2017).

In the genetics community, GWAS summary association statistics (e.g., effect size, standard error, p-value) of all SNPs for various traits are shared and assembled into large databases. Summary statistics from more than 4000 GWAS are now publicly available (Watanabe et al., 2018) and the number rises steeply. As individual-level SNP data are massive and are often under strict ethical/regulatory protections, it is an active research area to directly use these GWAS summary statistics for various in-sample and out-of-sample analyses (Pasaniuc and Price, 2017). For example, GWAS summary statistics are used to prioritize causal variants in fine-mapping analysis (Schaid et al., 2018), to quantify genetic overlaps among different traits (Bulik-Sullivan et al., 2015; Speed and Balding, 2019), to perform causal inference among traits via Mendelian randomization (Zhao et al., 2018), and to carry out integrative association tests with gene expression data (Gamazon et al., 2015; Gusev et al., 2016; Hu et al., 2019).

2.1.1 Heritability and genetic correlation

In GWAS, genetics signal strength and their genetics overlaps are often quantified as heritability and genetic correlation, and are standard measures to report. Many statistical methods have been developed on the use of common SNP data to infer the heritability and cross-trait genetic correlation in general populations. For instance, heritability h^2 can be estimated by aggregating the small contributions of a large number of common SNP markers, resulting in the SNP heritability estimator (Yang et al., 2010, 2017; Loh et al., 2015). Moreover, genetic correlation quantifies the genetic relationship between two heritable phenotypes and is traditionally estimated in family studies. GWAS data offer an alternative to family studies for genetic correlation estimation using independent individuals (van Rheenen et al., 2019). Specifically, GWAS data are able to measure the genetic similarity attributable to common SNPs, which can be calculated as the Pearson correlation of the genetic effects of SNPs on the two traits (Bulik-Sullivan et al., 2015; Speed and Balding, 2019).

Statistically, Jiang et al. (2016) shows that the REML estimator of heritability is consistent in high-dimensional LMM regardless of the number of causal SNPs, and this estimator (named GREML) has been implemented in the popular genetic tool GCTA (Yang et al., 2011). The theoretical results on heritability in Jiang et al. (2016) are built on the assumption that SNPs are independent. The GREML estimator might be biased when SNPs are correlated (Ma and Dicker, 2019) or hidden confounding effects (Holmes et al., 2019), though such bias is often believed to be small and acceptable in practice. See Yang et al. (2017) and van Rheenen et al. (2019) for overviews of these genetic concepts, and a detailed numerical comparison of population methods in Evans et al. (2018).

2.1.2 Complex trait prediction

Owing primarily to the potential to translate GWAS findings to medical advancements, it is of particular interest to predict the personalized genetic risk for new GWAS individuals using results from historical GWAS (Torkamani et al., 2018; Sugrue and Desikan, 2019; Martin et al., 2019). One of the state-of-art methods for genetic risk prediction of human

complex traits is genome-wide polygenic risk score (PRS) (Purcell et al., 2009), which is a weighted sum of millions of SNPs where each SNP is weighted by their estimated effect size from discovery GWAS. As no need to access the personal DNA information of subjects in the training set, PRS is computationally efficient and has widespread applications with more than $3,000$ related publications in 2018 (Zhao and Zou, 2019). Recent efforts have begun to explore the clinical utility of PRS on human diseases, such as heart disease and breast cancer (Mavaddat et al., 2019; Khera et al., 2018).

Though GWAS summary statistics have numerous applications, there is little rigorous theoretical evaluation. Without sufficient understanding of their statistical properties, we risk drawing erroneous decisions on summary statistics database design and construction. Most, if not all, of publicly shared GWAS summary statistics are marginal effects generated from marginal screening. Let $\boldsymbol{y}$ be an $n \times 1$ vector of continuous trait, a linear polygenic structure between $\boldsymbol{y}$ and SNP data $\boldsymbol{X}$ is often assumed in GWAS (e.g., Jiang et al. (2016))

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon} = \sum_{i=1}^{m} \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m, \boldsymbol{x}_{m+1}, \cdots, \boldsymbol{x}_p)$ is an $n \times p$ SNP data matrix with population-level correlation $\boldsymbol{\Sigma}$ among the p features, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m, \beta_{m+1}, \cdots, \beta_p)^T$ is a $p \times 1$ vector of genetic effects such that $(\beta_1, \cdots, \beta_m)^T$ are m unknown nonzero parameters, and $(\beta_{m+1}, \cdots, \beta_p)^T$ are zeros, and the $n \times 1$ vector $\boldsymbol{\epsilon}$ represents independent non-genetic random errors. The single SNP analysis in GWAS is given by

$$\boldsymbol{y} = \mathbf{1}_{\boldsymbol{n}} \mu_i + \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon}_i^* \tag{2.1}$$

for $i = 1, \cdots, p$, which is a marginal screening approach similar to sure independence screening (Fan and Lv, 2008). Let $\widehat{\boldsymbol{\beta}}_S = (\widehat{\beta}_{1S}, \cdots, \widehat{\beta}_{pS})^T$ be the marginal screening ordinary least

squares (OLS) estimator of model (2.1), the marginal estimators are given by

$$\widehat{\beta}_{iS} = (\boldsymbol{x}_i^T \boldsymbol{x}_i)^{-1} \boldsymbol{x}_i^T \boldsymbol{y}, \quad i = 1, \cdots, p,$$

and thus $\widehat{\boldsymbol{\beta}}_S = \{\text{Diag}(\boldsymbol{X}^T\boldsymbol{X})\}^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is the form of nearly all shared GWAS summary statistics for continuous traits, where $\text{Diag}(\boldsymbol{A})$ is the diagonal of matrix $\boldsymbol{A}$.

For human complex traits, the number of causal SNPs m is trait-specific and population-specific, and it can be comparable with n, but is not necessarily p. An overwhelming number of empirical evidence supports the polygenicity and pleiotropy of complex traits (e.g., Watanabe et al. (2018); Martin et al. (2019); Sullivan and Geschwind (2019)), which can be potentially explained by biological complexity and negative selection (O'Connor et al., 2019). Statistically, we may have a dense (sparsity free) signal model, while not every feature has a nonzero effect on the outcome. This is different from the standard settings in spare regression (e.g., Zhao and Yu (2006); Fan and Lv (2008); Feng and Zhang (2017); Guo et al. (2019)), which often has sparsity restriction on m.

2.2 Ridge-type estimators

In this section, we summarize the estimators investigated in our analysis and highlight their natural connections.

2.2.1 Ridge, marginal, and ridge-less

For simplicity, suppose $\boldsymbol{X}$ have been column-standardized to have mean zero and variance one, then the marginal estimator can be asymptotically given by

$$\widehat{\boldsymbol{\beta}}_S = \{\text{Diag}(\boldsymbol{X}^T\boldsymbol{X})\}^{-1}\boldsymbol{X}^T\boldsymbol{y} = \{\text{Diag}(\widehat{\boldsymbol{\Sigma}}_X)\}^{-1} \cdot n^{-1}\boldsymbol{X}^T\boldsymbol{y} = n^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{2.2}$$

where $\widehat{\boldsymbol{\Sigma}}_X = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ is the sample covariance matrix. The ridge-regularized estimator (Hoerl and Kennard, 1970; Tikhonov, 1963) with regularization parameter λ is

$$\widehat{\boldsymbol{\beta}}_R(\lambda) = \big(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p\big)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \big(\widehat{\boldsymbol{\Sigma}}_X + \lambda\boldsymbol{I}_p\big)^{-1}\widehat{\boldsymbol{\beta}}_S, \quad \lambda \in (0,\infty). \tag{2.3}$$

Here $\widehat{\boldsymbol{\Sigma}}_X + \lambda\boldsymbol{I}_p$ is a linear combination of $\widehat{\boldsymbol{\Sigma}}_X$ and diagonal matrix $\lambda\boldsymbol{I}_p$, and is called linear shrinkage estimator of $\boldsymbol{\Sigma}$ (Ledoit and Wolf, 2004). In equation (2.3), $\widehat{\boldsymbol{\beta}}_R(\lambda)$ can be viewed as the marginal estimator $\widehat{\boldsymbol{\beta}}_S$ after "accounting for $\boldsymbol{\Sigma}$" through this linear shrinkage estimator. When λ is large enough such that $\lambda\boldsymbol{I}_p$ can dominate $\widehat{\boldsymbol{\Sigma}}_X$, $\widehat{\boldsymbol{\Sigma}}_X + \lambda\boldsymbol{I}_p$ becomes asymptotically a diagonal matrix. Thus, let $\widehat{\boldsymbol{\beta}}_R(\infty) = \lim_{\lambda\to\infty}\widehat{\boldsymbol{\beta}}_R(\lambda)$, as $\lambda \to \infty$, we can have

$$\widehat{\boldsymbol{\beta}}_R(\infty) \propto \widehat{\boldsymbol{\beta}}_S.$$

On the other hand, as $\lambda \to 0^+$ (from the right), we have the ridge-less least squares estimator (Hastie et al., 2019)

$$\widehat{\boldsymbol{\beta}}_R(0^+) = \lim_{\lambda\to 0^+}\widehat{\boldsymbol{\beta}}_R(\lambda) = \big(\boldsymbol{X}^T\boldsymbol{X}\big)^+\boldsymbol{X}^T\boldsymbol{y} = \widehat{\boldsymbol{\Sigma}}_X^+\widehat{\boldsymbol{\beta}}_S,$$

where $\boldsymbol{A}^+$ is the Moore-Penrose pseudoinverse of matrix $\boldsymbol{A}$. When $n > p$ and suppose $\boldsymbol{X}$ has full column rank, $\widehat{\boldsymbol{\beta}}_R(0^+)$ reduces to the classic OLS estimator $\widehat{\boldsymbol{\beta}}_O$ given by

$$\widehat{\boldsymbol{\beta}}_O = \widehat{\boldsymbol{\beta}}_R(0) = \big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \widehat{\boldsymbol{\Sigma}}_X^{-1}\widehat{\boldsymbol{\beta}}_S.$$

2.2.2 BLUP and BLUP-less

In addition, ridge estimators have natural connection with the following best linear unbiased prediction (BLUP)

$$\widehat{\boldsymbol{\beta}}_B(\tau) = \boldsymbol{X}^T\big(\boldsymbol{X}\boldsymbol{X}^T + \tau p\boldsymbol{I}_n\big)^{-1}\boldsymbol{y}, \quad \tau \in (0,\infty).$$

BLUP is originally from linear mixed effects model (LMM) (Henderson, 1975, 1950) and has been widely applied in genetics to tackle dense genetic effects (e.g., Yang et al. (2010)). Similar to $\widehat{\boldsymbol{\beta}}_R(0^+)$, we can define the BLUP-less estimator $\widehat{\boldsymbol{\beta}}_B(0^+)$ by letting $\tau \to 0$

$$\widehat{\boldsymbol{\beta}}_B(0^+) = \lim_{\tau \to 0^+} \widehat{\boldsymbol{\beta}}_B(\tau) = \boldsymbol{X}^T \big(\boldsymbol{X}\boldsymbol{X}^T\big)^+ \boldsymbol{y}.$$

When $n < p$ and suppose $\boldsymbol{X}$ has full row rank, $\boldsymbol{X}^T \big(\boldsymbol{X}\boldsymbol{X}^T\big)^+$ reduces to $\boldsymbol{X}^T \big(\boldsymbol{X}\boldsymbol{X}^T\big)^{-1}$, which has been used for variable selection (Wang and Leng, 2016) and also has many applications in genetics, such as OmicKriging (Wheeler et al., 2014). It can be shown that (Wang and Leng, 2016)

$$\boldsymbol{X}^T \big(\boldsymbol{X}\boldsymbol{X}^T + \tau p \boldsymbol{I}_n\big)^{-1} \boldsymbol{y} = \big(\boldsymbol{X}^T\boldsymbol{X} + \tau p \boldsymbol{I}_p\big)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

It follows that $\widehat{\boldsymbol{\beta}}_B(\tau) = \widehat{\boldsymbol{\beta}}_R(\tau\omega)$ and there is one-to-one correspondence between ridge estimator and BLUP.

Similar to $\widehat{\boldsymbol{\beta}}_S$, all of the ridge-type *conditional* estimators $\widehat{\boldsymbol{\beta}}_R(\lambda)$, $\widehat{\boldsymbol{\beta}}_B(\tau)$, $\widehat{\boldsymbol{\beta}}_R(0^+)$, $\widehat{\boldsymbol{\beta}}_B(0^+)$, and $\widehat{\boldsymbol{\beta}}_O$ can be shared as GWAS summary-level data for following-up in-sample and out-of-sample applications. However, their computational complexity can be totally different in large training dataset with both n and $p \to \infty$. Particularly, marginal estimator $\widehat{\boldsymbol{\beta}}_S$ is usually much less computationally expensive than conditional estimators. Thus, understanding their connections and differences are important to determine the GWAS summary-level data to share while considering the computation-accuracy trade-off. In the rest of this paper, we analyze and compare these estimators in an unifying framework. We name them the class of ridge-type estimators.

2.2.3 Related work

Our analysis is related to literature on the studies of high-dimensional linear model without sparsity assumption, most of which are on the asymptotic behavior of high-dimensional ridge

estimator, including Dicker (2013, 2016); Dobriban and Wager (2018); El Karoui (2013, 2018); Hastie et al. (2019); Hsu et al. (2011) and Pluta et al. (2017). For example, Dicker (2013, 2016) studies dense signal ridge problems with Gaussian assumption of data and allows general correlation structure $\mathbf{\Sigma}$ among predictors. Dobriban and Wager (2018) study ridge estimator without Gaussian assumption and recently extend their results to distributed computing problem (Dobriban and Sheng, 2019). El Karoui (2013, 2018) studies ridge estimator in robust regression. Motivated by interpolation in machine learning, Hastie et al. (2019) analyze the ridge-less estimator by taking a limit on regularization parameter λ. In addition, our results for $\omega \in (0, 1)$ are related to studies of OLS estimator in moderate-dimensions (Guo and Cheng, 2018; Yang and Cheng, 2018).

Our analysis is also related to previous studies on high-dimensional LMM (Jiang et al., 2016; Steinsaltz et al., 2018; Dicker and Erdogdu, 2017; Ma and Dicker, 2019), in which the authors mainly focus on the in-sample inference of LMM model parameters (such as h^2) and do not pay attention to BLUP and out-of-sample predictions. On the other hand, BLUP has been a popular method in genetics and agriculture for a long time (Robinson, 1991). Thus, there are studies of BLUP in genetics community, sometimes named genomic BLUP (gBLUP), such as Goddard (2009); Daetwyler et al. (2010); de los Campos et al. (2013); Speed and Balding (2014), and some Bayesian or ridge alternatives, such as Zhou et al. (2013) and Li et al. (2014).

Different from the above literature, we are particularly motivated by the increasing applications of marginal estimator in the high-dimensional dense setting, especially in out-of-sample genetic risk prediction of GWAS. On this topic, a few studies in genetics field (such as Daetwyler et al. (2008); Dudbridge (2013), and Zhao and Zou (2019)) have explored the special case $\mathbf{\Sigma} = \boldsymbol{I}_p$. To the best of our knowledge, there is no rigorous study on behaviors of marginal estimator in high-dimensional dense setting with general $\mathbf{\Sigma}$. In our analysis, we build our analysis on random matrix theory (RMT) and allow an arbitrary correlation structure $\mathbf{\Sigma}$ among features. Moreover, we link marginal estimator to ridge estimator and

BLUP, and compare them in a unified framework.

Furthermore, we focus on R-squared (R^2) instead of the mean squared prediction error (MSE), which is commonly studied in most of previous theoretical studies on ridge-type estimators. The $R^2 \in [0, 1]$ can be viewed as a normalized version of MSE. In addition, different from MSE, R^2 is invariant to linear transformations of predictors, such as scaling and adding constants. This enables us to quantify and compare the performance of different estimators in a unified manner. More importantly, using R^2 allows us to generalize our analysis to study the cross-trait prediction particularly when traits in training and testing data are different or have some heterogeneity. From a practical perspective, R^2 and pseudo R^2 are standard measures used in GWAS (and many other areas) for evaluating the out-of-sample prediction performance and in-sample goodness-of-fit. Thus, our results on R^2 can be easily applied in practice, such as in sample size calculation for a desired prediction accuracy goal of GWAS.

2.2.4 Assembled estimators

Assembled estimators can be efficiently generated by combining together the marginal estimator learned from training data and the feature covariance structure estimated on an independent reference panel. In GWAS complex trait prediction, individual-level training data $(\boldsymbol{X}, \boldsymbol{y})$ are often not accessible, and only summary-level data $\widehat{\boldsymbol{\beta}}_S$ are shared to public. To construct ridge estimator in such situations, a common practice is to estimate the $\boldsymbol{\Sigma}_X$ with external data. (Vilhjálmsson et al., 2015; Ge et al., 2019). The assembled ridge estimator can be defined as

$$\widehat{\boldsymbol{\beta}}_A(\lambda) = \left(\boldsymbol{W}^T\boldsymbol{W} + \lambda n_w \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \propto \left(\boldsymbol{W}^T\boldsymbol{W} + \lambda n_w \boldsymbol{I}_p\right)^{-1}\widehat{\boldsymbol{\beta}}_S, \quad \lambda \in (0, \infty), \tag{2.4}$$

where $\boldsymbol{W}$ is a $n_w \times p$ external SNP data matrix that is independent of $\boldsymbol{X}$. In $\widehat{\boldsymbol{\beta}}_A(\lambda)$, the variance-covariance structure $\boldsymbol{\Sigma}_X$ is estimated by $\widehat{\boldsymbol{\Sigma}}_A = n_w^{-1} \cdot \boldsymbol{W}^T\boldsymbol{W}$. We consider two different versions of $\widehat{\boldsymbol{\beta}}_A(\lambda)$ that are common in practice:

- $\boldsymbol{W}$ is from the testing data, donated as $\widehat{\boldsymbol{\beta}}_{A_1}(\lambda)$; and

- $\boldsymbol{W}$ is independent of both training and testing data, donated as $\widehat{\boldsymbol{\beta}}_{A_2}(\lambda)$. In other words, the $\boldsymbol{\Sigma}_X$ is estimated by external publicly available data (Tam et al., 2019), for example the 1000 Genomes Project LD reference panel (1000-Genomes-Project-Consortium., 2015).

We extend our analysis on ridge-type estimators to investigate the prediction performance of $\widehat{\boldsymbol{\beta}}_A(\lambda)$, and compare it with the original ridge estimator $\widehat{\boldsymbol{\beta}}_R(\lambda)$ and marginal estimator $\widehat{\boldsymbol{\beta}}_S$. In GWAS, the $\boldsymbol{\Sigma}_X$ of SNP data is known to have a block-diagonal structure (Pasaniuc and Price, 2017), which enables us to separately construct the assembled ridge estimator within each block. Therefore, we also evaluate the prediction performance of block-wise assembling. To handle the high collinearity among SNPs within the block, we propose to use block-wise local principal components (BLPCs) instead of the raw SNPs data to perform prediction. We show that BLPC-based assembled estimator can outperform SNP-based estimators on a wide range of complex traits in the UK Biobank database.

CHAPTER 3: GENETIC CORRELATION ESTIMATION IN GWAS

Accessing individual-level SNP data is often inconvenient due to policy restrictions, and a recent standard practice in the genetic community is to share the summary association statistics, including the estimated effect size, standard error, p-value, and sample size n, of all SNPs after GWAS are published. Therefore, it has became an active research area to examine the heritability and cross-trait genetic correlation based on GWAS summary statistics. Among them, the cross-trait polygenic risk score (PRS) (Purcell et al., 2009; Power et al., 2015) has became a popular routine to measure genetic similarity of polygenic traits with widespread applications (Hagenaars et al., 2016; Nivard et al., 2017; Socrates et al., 2017; Pouget et al., 2019). Compared with other popular methods such as the cross-trait linkage disequilibrium (LD) score regression (Bulik-Sullivan et al., 2015) (cross-trait LDSC), Bivariate GCTA (Lee et al., 2012), and BOLT-REML (Loh et al., 2015), cross-trait PRS offers at least two unique strengths as follows. First, cross-trait PRS only requires the GWAS summary statistics of one trait obtained from a large discovery GWAS, while it allows those of the other trait obtained from a much smaller testing GWAS dataset. In contrast, most other methods require large GWAS data for both traits on either summary or individual-level. Second, cross-trait PRS can provide genetic propensity for each sample in the testing dataset, enabling further prediction and treatment. However, given these strengths of cross-trait PRS, empirical evidence has shown a common bias phenomenon that even highly significant cross-trait PRS can only account for a very small amount of variance (R^2 can be $< 1\%$) when dissecting the shared genetic basis among highly related heritable traits (Clarke et al., 2016; Mistry et al., 2018; Bogdan et al., 2018). Except for some introductory studies, such as Daetwyler et al. (2008), Dudbridge (2013), and Visscher et al. (2014), few attempts have ever been made to rigorously study cross-trait PRS and to explain such a counterintuitive

phenomenon.

Here we fill this significant gap with the following contributions. By comprehensively investigating the properties of cross-trait PRS for polygenic/omnigenic traits, our first contribution is to show that the estimated genetic correlation may asymptotically biased towards zero, uncovering that the underlying genetic overlap can be seriously underestimated. Furthermore, when all p SNPs are used in cross-trait PRS, we show that the asymptotic bias is largely determined by the triple (n, p, h^2) and is independent of the unknown number of causal SNPs of the two traits. Thus, our second contribution is to propose a consistent estimator by correcting such asymptotic bias in cross-trait PRS. We also develop a novel estimator of genetic correlation which only requires two sets of summary statistics from large discovery GWAS. In addition, we study the variance of cross-trait PRS and explain why the estimator can still be significant when it is heavily biased towards zero.

Next, we show that when cross-trait PRS is constructed using q top-ranked SNPs whose GWAS p-values pass a given threshold, in addition to (n, p, h^2), the asymptotic bias will also be determined by the number of causal SNPs m, since the ratio m/n determines the quality of the q selected SNPs. Particularly, for highly polygenic/omnigenic traits with dense SNP signals, such screening may fail, resulting in larger bias in genetic correlation estimation. Based on these results, we provide practical guidelines for assessing the m/n ratio and minimizing the potential bias in real data applications. Finally, we generalize our results to quantify the influence of overlapping samples among GWAS. We show that our bias-corrected estimator for independent GWAS can be smoothly extended to GWAS with partially or even fully overlapping samples.

3.1 Cross-trait PRS with all SNPs

Since cross-trait PRS is designed for polygenic traits based on their GWAS summary statistics, we first introduce the polygenic model and highlight some properties of GWAS summary statistics. For common SNPs, the standard approach in GWAS is marginal screening.

That is, the marginal association between the phenotype and single SNP is assessed each at a time, while adjusting for the same set of covariates. Marginal screening procedures often work well to prioritize important variables given that the signals are sparse (Fan and Lv, 2008), but they may have noisy outcomes when signals are dense (Fan et al., 2012), which is often the case for GWAS of highly polygenic traits.

3.1.1 Polygenic trait

Let $\boldsymbol{X}_{(1)}$ be an $n \times m$ matrix of the SNP data with nonzero effects, and $\boldsymbol{X}_{(2)}$ be an $n \times (p-m)$ matrix of the null SNPs, resulting in an $n \times p$ matrix of all SNPs, donated by $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m, \boldsymbol{x}_{m+1}, \cdots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ is an $n \times 1$ vector of the SNP i, $i = 1, \cdots, p$. Columns of $\boldsymbol{X}$ are assumed to be independent after LD-based pruning. Further, we assume column-wise standardization on $\boldsymbol{X}$ is performed such that each variable has sample mean zero and sample variance one. Therefore, we may introduce the following condition on SNP data:

Condition 3.1. *Entries of $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}]$ are real-value independent random variables with mean zero, variance one and a finite eighth order moment.*

Let $\boldsymbol{y}$ be an $n \times 1$ vector of continuous polygenic phenotype. We assume a linear polygenic structure between $\boldsymbol{y}$ and $\boldsymbol{X}$ as follows:

$$\boldsymbol{y} = \sum_{i=1}^{p} \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon} = \sum_{i=1}^{m} \boldsymbol{x}_i \beta_i + \boldsymbol{\epsilon} = \boldsymbol{X}_{(1)} \boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{\beta}^T = (\beta_1, \cdots, \beta_m, \beta_{m+1}, \cdots, \beta_p) = \left(\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T\right)$ is an vector of parameters such that β_i in $\boldsymbol{\beta}_{(1)} = (\beta_1, \cdots, \beta_m)^T$ are unknown nonzero genetic effects ($i = 1, \cdots, m$), and $\boldsymbol{\beta}_{(2)} = (\beta_{m+1}, \cdots, \beta_p)^T$ are zeros. $\boldsymbol{\epsilon}$ represents the vector of independent non-genetic random errors. Since the distribution assumption of $\boldsymbol{\beta}$ is not necessary to illustrate GWAS marginal screening below, we simply treat $\boldsymbol{\beta}_{(1)}$ as fixed unknown parameters in this subsection. We will introduce detailed distribution assumption on $\boldsymbol{\beta}$ for cross-trait PRS analysis in Section 3.1.2. For simplicity, we assume that there are no other fixed effects in model (3.1), or equivalently,

other covariates can be well observed and adjusted for.

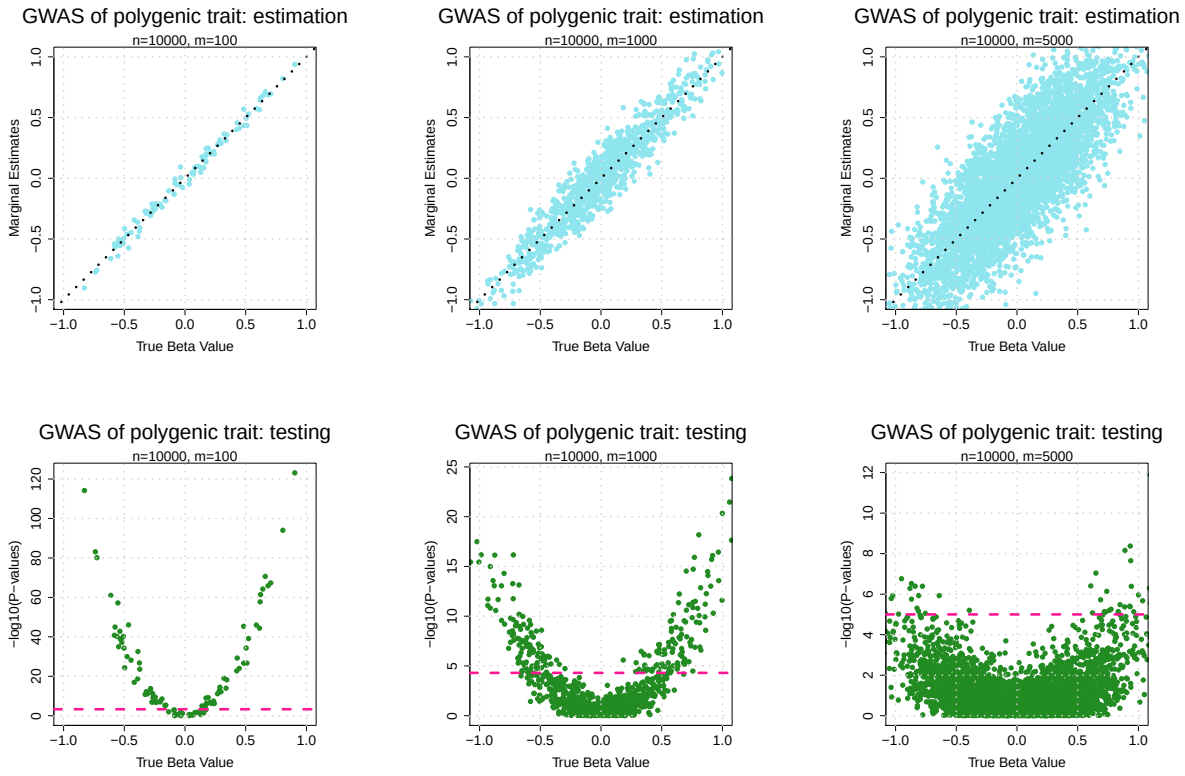

Figure 3.1: Estimation (upper panels) and testing (bottom panels) of marginal genetic effects in GWAS of polygenic traits. We set $n = 10,000$ and $m = p = 100$, 1000 and 5000.

The m/n ratio is important for genetic effects estimation under polygenic model (3.1). Intuitively, m represents the dispersion of nonzero genetic effects, and $1/m$ represents the relative per-SNP contribution to $\boldsymbol{y}$ (i.e., signal strength). Since $1/n$ quantifies the sample error in estimating genetic effects, m and n will compete against each other in genetic effects estimation. For GWAS marginal screening, such competition can be easily illustrated by the mean and variance of genetic effect estimates. Consider a GWAS that performs single SNP analysis given by

$$\boldsymbol{y} = \mathbf{1}_{\boldsymbol{n}}\mu_i + \boldsymbol{x}_i\beta_i + \boldsymbol{\epsilon}_i^* \tag{3.2}$$

for $i = 1, \cdots, p$, where $\mathbf{1}_{\boldsymbol{n}}$ is an $n \times 1$ vector of ones. Let $\widehat{\mu}_i$ and $\widehat{\beta}_i$ be the ordinary least

squares (OLS) estimates of μ_i and β_i in marginal screening model (3.2), respectively, for $i = 1, \cdots, p$. When both n and $m \to \infty$, under Condition 3.1 and model (3.1), it can be shown that

$$\mathrm{E}\big(\widehat{\mu}_i\big) = \mu_i, \qquad \mathrm{E}\big(\widehat{\beta}_i\big) = \beta_i, \qquad \text{and}$$

$$\mathrm{Var}\big(\widehat{\beta}_i\big) = \begin{cases} n^{-1} \cdot \big(\sum_{j\neq i}^{m} \beta_j^2 + 1\big), & \text{for} \quad i \in [1, m]; \\ n^{-1} \cdot \big(\sum_{j=1}^{m} \beta_j^2 + 1\big), & \text{for} \quad i \in [m+1, p]. \end{cases} \tag{3.3}$$

It follows that the T scores for testing

$$H_{0i} : \beta_i = 0 \qquad \text{versus} \qquad H_{1i} : \beta_i \neq 0, \quad \text{for} \quad i = 1, \cdots, p$$

are given by

$$T_i = \begin{cases} \widehat{\beta}_i / \{(\sum_{j\neq i}^{m} \beta_j^2 + 1)/n\}^{1/2} = \widehat{\beta}_i \cdot \sqrt{n/(\sum_{j\neq i}^{m} \beta_j^2 + 1)}, & \text{for} \quad i \in [1, m]; \\ \widehat{\beta}_i / \{(\sum_{j=1}^{m} \beta_j^2 + 1)/n\}^{1/2} = \widehat{\beta}_i \cdot \sqrt{n/(\sum_{j=1}^{m} \beta_j^2 + 1)}, & \text{for} \quad i \in [m+1, p] \end{cases}$$

under H_{0i}, $i = 1, \cdots, p$.

Remark 3.1. *The term $\sum_{j=1}^{m} \beta_j^2$ (or $\sum_{j\neq i}^{m} \beta_j^2$) in Equation (3.3) is induced by the cumulative spurious correlations (Fan et al., 2018) between the SNP i and the m (or $m-1$) causal SNPs. When m/n is large, $E(\widehat{\beta}_i) = \beta_i$ might not dominate the standard error ($\{(\sum_{j\neq i}^{m} \beta_j^2 + 1)/n\}^{1/2}$, for $i = 1, \cdots, m$; or $\{(\sum_{j=1}^{m} \beta_j^2 + 1)/n\}^{1/2}$, for $i = m+1, \cdots, p$). More importantly, all standard error are in the same scale regardless of whether their original β_is are zeros or not. Thus, the $\widehat{\beta}_i$s from causal and null variants can be totally mixed up when m/n is large. Then the test statistics T_is may not well preserve the ranking of variables in $\boldsymbol{X}$ when m/n is large, resulting in potential low power and high false positive rate in detecting and prioritizing important SNPs using GWAS marginal screening p-values.*

Figure 3.1 demonstrates the estimation and testing of marginal genetic effects in GWAS with $n = 10,000$ as $p = m$ increases from 100, 1000 to 5000. Each entry of $\boldsymbol{X}$ is i.i.d generated from $N(0,1)$, elements of $\boldsymbol{\beta}_{(1)}$ are i.i.d generated from $N(0,0.4)$, and entries of $\boldsymbol{\epsilon}$ are i.i.d from $N(0,1)$. Then, $\boldsymbol{y}$ is generated from model (3.1). The estimated genetic effects are unbiased in general, however, the uncertainty clearly moves up as m increases. The relative contribution of each SNP decreases as m increases, and thus the testing power drops as well. More simulations on GWAS summary statistics can be found in Section 3.4.

As illustrated in later sections, the asymptotic behavior of cross-trait PRS and the performance of SNP screening are closely related to the ratios among (n, m, p). Specifically, 1) when cross-trait PRS is constructed with all p SNPs, the sample errors of the p $\widehat{\beta}_i$s are aggregated, resulting in inflated genetic variance and underestimated genetic correlation; and ii) when cross-trait PRS is constructed with top-ranked SNPs that pass a pre-specified p-value threshold, it may have worse performance if GWAS marginal screening fails to prioritize the causal SNPs.

3.1.2 General setup

We first introduce the modelling framework to investigate the cross-trait PRS, including the genetic architecture of polygenic traits, distribution of genetic effects, and genetic correlation estimators. Consider three independent GWAS that are conducted for three different traits as follows:

- Discovery GWAS-I: $(\boldsymbol{X}, \boldsymbol{y}_\alpha)$, with $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{X}_{(1)} \in \mathbb{R}^{n_1 \times m_\alpha}$, and $\boldsymbol{y}_\alpha \in \mathbb{R}^{n_1 \times 1}$.
- Discovery GWAS-II: $(\boldsymbol{Z}, \boldsymbol{y}_\beta)$, with $\boldsymbol{Z} = [\boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)}] \in \mathbb{R}^{n_2 \times p}$, $\boldsymbol{Z}_{(1)} \in \mathbb{R}^{n_2 \times m_\beta}$, and $\boldsymbol{y}_\beta \in \mathbb{R}^{n_2 \times 1}$.
- Target testing GWAS: $(\boldsymbol{W}, \boldsymbol{y}_\eta)$, with $\boldsymbol{W} = [\boldsymbol{W}_{(1)}, \boldsymbol{W}_{(2)}] \in \mathbb{R}^{n_3 \times p}$, $\boldsymbol{W}_{(1)} \in \mathbb{R}^{n_3 \times m_\eta}$, and $\boldsymbol{y}_\eta \in \mathbb{R}^{n_3 \times 1}$.

Here $\boldsymbol{y}_\alpha$, $\boldsymbol{y}_\beta$, and $\boldsymbol{y}_\eta$ are three different continuous phenotypes studied in three GWAS with

sample sizes n_1, n_2, and n_3, respectively. The m_α, m_β, and m_η are different numbers of causal SNPs in general. The $\boldsymbol{X}_{(1)}$, $\boldsymbol{Z}_{(1)}$, and $\boldsymbol{W}_{(1)}$ denote the causal SNPs of $\boldsymbol{y}_\alpha$, $\boldsymbol{y}_\beta$, and $\boldsymbol{y}_\eta$, respectively, and $\boldsymbol{X}_{(2)}$, $\boldsymbol{Z}_{(2)}$, and $\boldsymbol{W}_{(2)}$ donate the corresponding null SNPs. Thus, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are three matrices of p SNPs. Further, we assume column-wise standardization on $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ is performed such that each variable has sample mean zero and sample variance one. Therefore, we may introduce the following condition on SNP data:

Condition 3.2. *As* $min(n, p) \to \infty$, *we assume*

$$\frac{m}{n} = \gamma \to \gamma_0 \quad \textit{and} \quad \frac{m}{p} = \omega \to \omega_0 \quad \textit{for} \quad 0 < \gamma_0 \leq \infty \quad \textit{and} \quad 0 \leq \omega_0 \leq 1,$$

which should satisfy most large-scale GWAS of polygenic traits.

It is assumed that $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ have been standardized and satisfy Condition 3.1. Similar to model (3.1), we assume a linear polygenic structure as follows:

$$\boldsymbol{y}_\alpha = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_\alpha, \quad \boldsymbol{y}_\beta = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}_\beta, \quad \text{and} \quad \boldsymbol{y}_\eta = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}_\eta, \tag{3.4}$$

where $\boldsymbol{\alpha}^T = \left(\boldsymbol{\alpha}_{(1)}^T, \boldsymbol{\alpha}_{(2)}^T\right)$, $\boldsymbol{\beta}^T = \left(\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T\right)$, and $\boldsymbol{\eta}^T = \left(\boldsymbol{\eta}_{(1)}^T, \boldsymbol{\eta}_{(2)}^T\right)$ are $p \times 1$ vectors of SNP effects, in which $\boldsymbol{\alpha}_{(2)}$, $\boldsymbol{\beta}_{(2)}$, and $\boldsymbol{\eta}_{(2)}$ are zeros, and $\boldsymbol{\epsilon}_\alpha$, $\boldsymbol{\epsilon}_\beta$, and $\boldsymbol{\epsilon}_\eta$ represent independent random error vectors. The $\boldsymbol{\alpha}_{(1)}$, $\boldsymbol{\beta}_{(1)}$, and $\boldsymbol{\eta}_{(1)}$ are random variables (Dobriban and Wager, 2018; Jiang et al., 2016), and the distribution assumption is detailed in the following subsection. The overall genetic heritability of $\boldsymbol{y}_\alpha$ is, therefore, given by $h_\alpha^2 = \text{Var}(\boldsymbol{X}\boldsymbol{\alpha})/\text{Var}(\boldsymbol{y}_\alpha) = \text{Var}(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)})/\{\text{Var}(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)}) + \text{Var}(\boldsymbol{\epsilon}_{\boldsymbol{\alpha}})\}$, which measures the proportion of variation in $\boldsymbol{y}_\alpha$ that can be explained by the aggregated genetic variation $\text{Var}(\boldsymbol{X}\boldsymbol{\alpha})$. The $\boldsymbol{y}_\alpha$ is fully heritable when $h_\alpha^2 = 1$. Similarly, we can define the heritability h_β^2 of $\boldsymbol{y}_\beta$ and h_η^2 of $\boldsymbol{y}_\eta$, respectively. We assume h_α^2, h_β^2, and $h_\eta^2 \in (0, 1]$. The genetic correlation in this chapter is defined as the correlation of SNP effects on pairs of phenotypes (Lu et al., 2017; Pasaniuc and Price, 2017; Shi et al., 2017; Guo et al., 2019).

Definition 3.1 (Genetic Correlation)**.** *The genetic correlation between* $\boldsymbol{y}_\alpha$ *and* $\boldsymbol{y}_\eta$ *and that between* $\boldsymbol{y}_\alpha$ *and* $\boldsymbol{y}_\beta$ *are respectively given by*

$$\varphi_{\alpha\eta} = \frac{\boldsymbol{\alpha}^T\boldsymbol{\eta}}{\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\|} \cdot I(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\| > 0) \quad \textit{and} \quad \varphi_{\alpha\beta} = \frac{\boldsymbol{\alpha}^T\boldsymbol{\beta}}{\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\beta}\|} \cdot I(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\beta}\| > 0),$$

where $I(\cdot)$ *is the indicator function,* $\|\cdot\|$ *is the* l_2 *norm of a vector, and* $\varphi_{\alpha\eta}$ *and* $\varphi_{\alpha\beta} \in [-1, 1]$.

Genetic effects In this subsection, we introduce the distribution assumption on nonzero genetic effects $\boldsymbol{\alpha}_{(1)}$, $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\eta}_{(1)}$. Since m_α, m_β and m_η can be different and the causal SNPs of different phenotypes may partially overlap, we let $m_{\alpha\eta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$, and $m_{\alpha\beta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$. Let $F(0, V)$ represent a generic distribution with mean zero, (co)variance V, and finite fourth order moments. We introduce the following condition on genetic effects and random errors.

Condition 3.3. *As* n_1, n_3, $p \to \infty$, $m_{\alpha\eta}, m_\alpha$, *and* $m_\eta \to \infty$, *we assume* $m_{\alpha\eta}/\sqrt{m_\alpha m_\eta} = \kappa_{\alpha\eta} \to \kappa_{0\alpha\eta} \in (0, 1]$. *Similarly, as* n_1, n_2, n_3, $p \to \infty$, $m_{\alpha\beta}, m_\alpha$, *and* $m_\beta \to \infty$, *we assume* $m_{\alpha\beta}/\sqrt{m_\alpha m_\beta} = \kappa_{\alpha\beta} \to \kappa_{0\alpha\beta} \in (0, 1]$. α_i, β_j, *and* η_k *are independent random variables satisfying*

$$\alpha_i \sim F(0, \sigma_\alpha^2/p), \quad i = 1, ..., m_\alpha; \qquad \beta_j \sim F(0, \sigma_\beta^2/p), \quad j = 1, ..., m_\beta;$$
$$\eta_k \sim F(0, \sigma_\eta^2/p), \quad k = 1, ..., m_\eta.$$

The $m_{\alpha\eta}$ *overlapping nonzero effects* $(\alpha_i, \eta_i)s$ *of* ($\boldsymbol{y}_\alpha$,$\boldsymbol{y}_\eta$) *and* $m_{\alpha\beta}$ *overlapping nonzero effects* $(\alpha_j, \beta_j)s$ *of* ($\boldsymbol{y}_\alpha$,$\boldsymbol{y}_\beta$) *satisfy*

$$\begin{pmatrix} \alpha_i \\ \eta_i \end{pmatrix} \sim F\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, p^{-1} \cdot \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\eta} \\ \sigma_{\alpha\eta} & \sigma_\eta^2 \end{pmatrix}\right] \quad \textit{and} \quad \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim F\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, p^{-1} \cdot \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}\right],$$

respectively. And ϵ_{α_i}, ϵ_{β_j} *and* ϵ_{η_k} *are independent random variables satisfying*

$$\epsilon_{\alpha i} \sim F(0, \sigma^2_{\epsilon_\alpha}), \quad i = 1, ..., n_1; \qquad \epsilon_{\beta j} \sim F(0, \sigma^2_{\epsilon_\beta}), \quad j = 1, ..., n_2;$$
$$\epsilon_{\eta k} \sim F(0, \sigma^2_{\epsilon_\eta}), \quad k = 1, ..., n_3;$$

where $\sigma_{\alpha\eta} = \rho_{\alpha\eta} \cdot \sigma_\alpha \sigma_\eta$ *and* $\sigma_{\alpha\beta} = \rho_{\alpha\beta} \cdot \sigma_\alpha \sigma_\beta$.

Since the three GWAS have independent samples, we assume that their random errors are independent. Overlapping samples and the induced non-genetic correlation will be studied in Section 3.3. The genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ is asymptotically given by

$$\varphi_{\alpha\eta} = \frac{\boldsymbol{\alpha}^T \boldsymbol{\eta}}{\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\|} = \frac{\sum_{i=1}^{m_{\alpha\eta}} \alpha_i \eta_i}{(\sum_{i=1}^{m_\alpha} \alpha_i^2)^{1/2} (\sum_{i=1}^{m_\eta} \eta_i^2)^{1/2}} = \kappa_{0\alpha\eta} \cdot \rho_{\alpha\eta} + o_p(1),$$

and the genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ is asymptotically given by

$$\varphi_{\alpha\beta} = \frac{\boldsymbol{\alpha}^T \boldsymbol{\beta}}{\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\beta}\|} = \frac{m_{\alpha\beta}}{(m_\alpha m_\beta)^{1/2}} \cdot \rho_{\alpha\beta} + o_p(1) = \kappa_{0\alpha\beta} \cdot \rho_{\alpha\beta} + o_p(1).$$

As in Jiang et al. (2016), heritability h^2_α, h^2_β, and h^2_η can be asymptotically represented as follows:

$$h^2_\alpha = \frac{(m_\alpha/p)\sigma^2_\alpha}{(m_\alpha/p)\sigma^2_\alpha + \sigma^2_{\epsilon_\alpha}}, \quad h^2_\beta = \frac{(m_\beta/p)\sigma^2_\beta}{(m_\beta/p)\sigma^2_\beta + \sigma^2_{\epsilon_\beta}}, \quad \text{and} \quad h^2_\eta = \frac{(m_\eta/p)\sigma^2_\eta}{(m_\eta/p)\sigma^2_\eta + \sigma^2_{\epsilon_\eta}}.$$

The aim of introducing the normalizer p^{-1} for nonzero genetic effects is to let the per-SNP contribution vanish and thus the aggregated genetic variation $\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta})$ remains finite (Bulik-Sullivan et al., 2015; Dobriban and Wager, 2018). It is also possible to introduce the normalization via SNP data as in Jiang et al. (2016). We note that the following analysis of cross-trait PRS remains the same in both situations, because the normalization will cancel out from the numerator and denominator of genetic correlation estimators.

Genetic correlation estimators For common SNPs, the standard approach in GWAS is marginal screening (Fan and Lv, 2008). That is, the marginal association between the phenotype and single SNP is assessed each at a time, while adjusting for the same set of covariates. Now we introduce the cross-trait PRS and genetic correlation estimators based on GWAS marginal screening. We need the following data. As n_1, n_2, and $p \to \infty$, the summary association statistics for $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ from Discovery GWAS-I & II are given by

$$\widehat{\boldsymbol{\alpha}} = \frac{1}{n_1}\boldsymbol{X}^T\boldsymbol{y}_\alpha = \frac{1}{n_1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \frac{1}{n_2}\boldsymbol{Z}^T\boldsymbol{y}_\beta = \frac{1}{n_2}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big).$$

We assume that the individual-level SNP $\boldsymbol{W}$ and phenotype $\boldsymbol{y}_\eta$ in the Target testing GWAS can be accessed. In addition, h^2_α, h^2_β, and h^2_η are assumed to be estimable, using either their corresponding individual-level data (Yang et al., 2010; Loh et al., 2015) or summary-level data (Bulik-Sullivan et al., 2015; Speed and Balding, 2019), or can be found in the literature (Polderman et al., 2015). In summary, besides (n_1, n_2, n_3, p), it is assumed that $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, $\boldsymbol{W}$, $\boldsymbol{y}_\eta$, $\widehat{h}^2_\alpha$, $\widehat{h}^2_\beta$, and $\widehat{h}^2_\eta$ are available. We construct cross-trait PRSs as follows:

$$\widehat{\boldsymbol{S}}_\alpha = \sum_{i=1}^{p} \boldsymbol{w}_i\widehat{a}_i = \boldsymbol{W}\widehat{\boldsymbol{a}} = \boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)} \quad \text{for } \boldsymbol{y}_\alpha \text{ and}$$
$$\widehat{\boldsymbol{S}}_{\boldsymbol{\beta}} = \sum_{i=1}^{p} \boldsymbol{w}_i\widehat{b}_i = \boldsymbol{W}\widehat{\boldsymbol{b}} = \boldsymbol{W}_{(1,\beta)}\widehat{\boldsymbol{b}}_{(1)} + \boldsymbol{W}_{(2,\beta)}\widehat{\boldsymbol{b}}_{(2)} \quad \text{for } \boldsymbol{y}_\beta,$$

where $\widehat{\boldsymbol{a}}^T = (\widehat{a}_1, \cdots, \widehat{a}_{m_\alpha}, \widehat{a}_{m_\alpha+1}, \cdots, \widehat{a}_p) = \big(\widehat{\boldsymbol{a}}_{(1)}^T, \widehat{\boldsymbol{a}}_{(2)}^T\big)$, in which $\widehat{a}_i = \widehat{\alpha}_i \cdot \mathrm{I}(|\widehat{\alpha}_i| > c_\alpha)$, $\widehat{\boldsymbol{b}}^T = (\widehat{b}_1, \cdots, \widehat{b}_{m_\beta}, \widehat{b}_{m_\beta+1}, \cdots, \widehat{b}_p) = \big(\widehat{\boldsymbol{b}}_{(1)}^T, \widehat{\boldsymbol{b}}_{(2)}^T\big)$, in which $\widehat{b}_i = \widehat{\beta}_i \cdot \mathrm{I}(|\widehat{\beta}_i| > c_\beta)$, and c_α and c_β are given thresholds used for SNP screening in order to calculate $\widehat{\boldsymbol{S}}_\alpha$ and $\widehat{\boldsymbol{S}}_\beta$. Moreover, we define $\boldsymbol{W}_{(1,\alpha)} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{m_\alpha}]$, $\boldsymbol{W}_{(2,\alpha)} = [\boldsymbol{w}_{m_\alpha+1}, \cdots, \boldsymbol{w}_p]$, $\boldsymbol{W}_{(1,\beta)} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{m_\beta}]$, $\boldsymbol{W}_{(2,\beta)} = [\boldsymbol{w}_{m_\beta+1}, \cdots, \boldsymbol{w}_p]$, and $\boldsymbol{W} = [\boldsymbol{W}_{(1,\alpha)}, \boldsymbol{W}_{(2,\alpha)}] = [\boldsymbol{W}_{(1,\beta)}, \boldsymbol{W}_{(2,\beta)}]$.

We estimate the genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ with $\big(\widehat{\boldsymbol{S}}_\alpha, \boldsymbol{y}_\eta\big)$ and that between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ with $\big(\widehat{\boldsymbol{S}}_\alpha, \widehat{\boldsymbol{S}}_\beta\big)$. They represent two common cases in real data applications. For $\big(\widehat{\boldsymbol{S}}_\alpha, \boldsymbol{y}_\eta\big)$, individual-level data are available for one trait, but not for another one. It often

occurs when the traits are studied in two different GWAS. For $\left(\widehat{\boldsymbol{S}}_{\alpha},\widehat{\boldsymbol{S}}_{\beta}\right)$, neither of the two traits has individual-level data. This happens when we have GWAS summary statistics of two traits and estimate their genetic correction on an independent target dataset. The genetic correlation estimators are given by

$$G_{\alpha\eta} = \frac{\boldsymbol{y}_{\eta}^{T}\widehat{\boldsymbol{S}}_{\alpha}}{\left\|\boldsymbol{y}_{\eta}\right\| \cdot \left\|\widehat{\boldsymbol{S}}_{\alpha}\right\|} = \frac{\left(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta}\right)^{T}\left(\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}\right)}{\left\|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta}\right\| \cdot \left\|\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}\right\|}$$

for $\varphi_{\alpha\eta}$, and

$$G_{\alpha\beta} = \frac{\widehat{\boldsymbol{S}}_{\beta}^{T}\widehat{\boldsymbol{S}}_{\alpha}}{\left\|\widehat{\boldsymbol{S}}_{\beta}\right\| \cdot \left\|\widehat{\boldsymbol{S}}_{\alpha}\right\|} = \frac{\left(\boldsymbol{W}_{(1,\beta)}\widehat{\boldsymbol{b}}_{(1)} + \boldsymbol{W}_{(2,\beta)}\widehat{\boldsymbol{b}}_{(2)}\right)^{T}\left(\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}\right)}{\left\|\boldsymbol{W}_{(1,\beta)}\widehat{\boldsymbol{b}}_{(1)} + \boldsymbol{W}_{(2,\beta)}\widehat{\boldsymbol{b}}_{(2)}\right\| \cdot \left\|\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}\right\|}$$

for $\varphi_{\alpha\beta}$.

3.1.3 Asymptotic bias and correction

We first investigate $G_{\alpha\beta}$ and $G_{\alpha\eta}$ when all of the p candidate SNPs are used, or when $c_{\alpha} = c_{\beta} = 0$. Thus, $\widehat{\boldsymbol{a}}_{(1)} = \widehat{\boldsymbol{\alpha}}_{(1)}$, $\widehat{\boldsymbol{a}}_{(2)} = \widehat{\boldsymbol{\alpha}}_{(2)}$, $\widehat{\boldsymbol{b}}_{(1)} = \widehat{\boldsymbol{\beta}}_{(1)}$, and $\widehat{\boldsymbol{b}}_{(2)} = \widehat{\boldsymbol{\beta}}_{(2)}$. Then, we have

$$G_{\alpha\eta} = \frac{\left(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta}\right)^{T}\boldsymbol{W}\boldsymbol{X}^{T}\left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha}\right)}{\left\|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta}\right\| \cdot \left\|\left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha}\right)^{T}\boldsymbol{X}\boldsymbol{W}^{T}\right\|}$$

and

$$G_{\alpha\beta} = \frac{\left(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta}\right)^{T}\boldsymbol{Z}\boldsymbol{W}^{T}\boldsymbol{W}\boldsymbol{X}^{T}\left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha}\right)}{\left\|\left(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\boldsymbol{\beta}}\right)^{T}\boldsymbol{Z}\boldsymbol{W}^{T}\right\| \cdot \left\|\left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha}\right)^{T}\boldsymbol{X}\boldsymbol{W}^{T}\right\|}.$$

We have the following results on the asymptotic properties of $G_{\alpha\eta}$, whose proof can be found in Appendix A.

Theorem 3.1. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose $m_{\alpha\eta}, m_{\alpha}$, and $m_{\eta} \to \infty$ as $min(n_1, n_3, p) \to \infty$, and let $p = c \cdot (n_1 n_3)^{a}$ for some constants $c > 0$ and*

$a \in (0, \infty]$. *If* $a \in (0, 1)$, *then we have*

$$G_{\alpha\eta} = \varphi_{\alpha\eta} + \left(\sqrt{\frac{n_1}{n_1 + p/h_\alpha^2}} \cdot h_\eta - 1 \right) \cdot \varphi_{\alpha\eta} + o_p(1).$$

If $a \in [1, \infty]$, *then we have* $G_{\alpha\eta} \cdot n_3 = O_p(1)$.

Remark 3.2. *The asymptotic limit of* $G_{\alpha\beta}$ *is independent of the unknown numbers* m_α, m_η, *and* $m_{\alpha\eta}$, *and is independent of the parameters of genetic effects in Condition 3.3. If* $a \in [1, \infty]$, *i.e.,* $p/(n_1 n_3)$ *is too large, then* $G_{\alpha\eta}$ *will have a zero asymptotic limit. In practice, this occurs when the sample size of discovery GWAS is too small to obtain reliable GWAS summary statistics. When these summary statistics are applied on an independent target dataset, the mean of genetic covariance* $\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_\alpha$ *cannot dominate its standard error. The genetic variance* $\widehat{\boldsymbol{S}}_\alpha^T \widehat{\boldsymbol{S}}_\alpha$ *is so overwhelming that* $G_{\alpha\eta}$ *goes to zero. Details can be found in Appendix A. If* $a \in (0, 1)$, $G_{\alpha\eta}$ *is a biased estimator of* $\varphi_{\alpha\eta}$ *when* $\sqrt{n_1/(n_1 + p/h_\alpha^2)} \cdot h_\eta$ *is smaller than* 1. *Formally, let* $p = c \cdot n_1^b$ *for some constants* $c > 0$ *and* $b \in (0, \infty]$, *we have*

$$\sqrt{\frac{n_1}{n_1 + p/h_\alpha^2}} \cdot h_\eta = \begin{cases} o_p(1), & \text{if} \quad b > 1; \\ \{h_\eta^2/(1 + c/h_\alpha^2)\}^{1/2}, & \text{if} \quad b = 1; \\ h_\eta, & \text{if} \quad b < 1. \end{cases}$$

It follows that $G_{\alpha\eta}$ *is an unbiased estimator of* $\varphi_{\alpha\eta}$ *only if* $h_\eta^2 = 1$ *and* $p = o(n_1)$. *For* $p = O(n_1)$, $G_{\alpha\eta}$ *is a shrinkage estimate of* $\varphi_{\alpha\eta}$; *and when* $n_1 = o(p)$, $G_{\alpha\eta}$ *is asymptotically zero. Therefore,* $G_{\alpha\eta}$ *has nonzero asymptotic limit only when training GWAS sample size* n_1 *is at least proportional to* p *(i.e.,* $b \leq 1$*). In such situation, a consistent estimator of* $\varphi_{\alpha\eta}$ *is given by*

$$G_{\alpha\eta}^A = G_{\alpha\eta} \cdot \sqrt{\frac{n_1 + p/h_\alpha^2}{n_1 \cdot h_\eta^2}} = \varphi_{\alpha\eta} + o_p(1).$$

The variance of $G_{\alpha\eta}$ and $G_{\alpha\eta}^A$ is provided in the following corollary.

Corollary 3.1. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose* $m_{\alpha\eta}, m_\alpha$, *and* $m_\eta \to \infty$ *as* $min(n_1, n_3, p) \to \infty$, *and let* $p = c \cdot n_1^b$ *for some constants* $c > 0$ *and* $b \in (0, 1]$, *we have*

$$Var(G_{\alpha\eta}) = \Big\{ \frac{(p + 2n_1 + 2n_3)h_\eta^2}{n_3(p/h_\alpha^2 + n_1)} \cdot \varphi_{\alpha\eta}^2 + \frac{n_1 h_\eta^2}{p/h_\alpha^2 + n_1} \cdot \frac{m_{\alpha\eta}(\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2)}{m_\alpha m_\eta \sigma_\alpha^2 \sigma_\eta^2} \Big\} \cdot \{1 + o_p(1)\},$$

where $E(\alpha_1^2 \eta_1^2) = \sigma_{\alpha^2\eta^2}/p^2$. *It follows that*

$$Var(G_{\alpha\eta}) = O_p\Big\{ \frac{n_1 + n_3}{n_3 n_1} + \frac{m_{\alpha\eta}}{m_\alpha m_\eta} \Big\} = O_p\Big\{ \max(\frac{1}{n_1}, \frac{1}{n_3}, \frac{1}{m_{\alpha\eta}}) \Big\}.$$

As the discovery GWAS sample size n_1 is often large, we usually have $n_1 > n_3$ in practice. Thus, Corollary 3.1 shows that the scale of $\mathrm{Var}(G_{\alpha\eta})$ is jointly determined by the testing GWAS sample size n_3 and the polygenicity of genetics co-architecture of the two traits, characterized by $m_{\alpha\eta}$. When $m_{\alpha\eta} \geq n_3$, $\mathrm{Var}(G_{\alpha\eta})$ has a scale $O_p(1/n_3)$ and thus the inference of $G_{\alpha\eta}$ can be valid in the testing GWAS even $G_{\alpha\eta}$ is heavily biased towards zero. For example, if $m_{\alpha\eta} \geq n_3$, the T score for testing $H_0 : \varphi_{\alpha\eta} = 0$ versus $H_1 : \varphi_{\alpha\eta} \neq 0$ is given by

$$T_{\alpha\eta}^2 = \frac{G_{\alpha\eta}^2}{\mathrm{Var}(G_{\alpha\eta})} = \Big\{ \frac{p + 2n_1 + 2n_3}{n_1 n_3} + \frac{\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2}{m_{\alpha\eta}\sigma_{\alpha\eta}^2} \Big\}^{-1} \cdot \{1 + o_p(1)\} = O_p(\frac{1}{n_3}),$$

under H_0. On the other hand, if $m_{\alpha\eta} < n_3$, cross-trait PRS may have large variance with scale $O_p(1/m_{\alpha\eta})$. Notably, the testing power of $G_{\alpha\eta}^A$ and $G_{\alpha\eta}$ is the same under the conditions of Corollary 3.1, because $G_{\alpha\eta}^A$ can be viewed as $G_{\alpha\eta}$ multiplies some constant.

In summary, estimating genetic correlation with cross-trait PRS requires the training GWAS sample size n_1 is at least proportional to p. The testing sample size n_3 vanishes in the limit of $G_{\alpha\eta}$, which verifies that we can apply the discovery summary statistics onto a much smaller set of target samples. In addition, the variance of cross-trait PRS have scale $O_p(1/n_3)$ for a pair of traits with high polygenicity (i.e., $m_{\alpha\eta} \geq n_3$). Therefore, cross-trait PRS may have good testing power even the estimation is biased. This result matches widespread

empirical observations that cross-trait PRS may have small p-value, but the R^2 is small. The asymptotic properties of $G_{\alpha\beta}$ are given as follows.

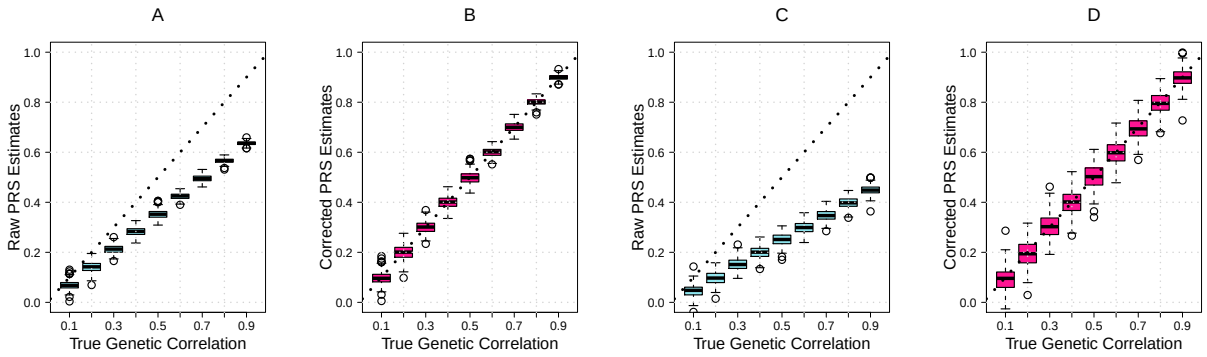

Figure 3.2: Raw genetic correlations estimated by cross-trait PRS with all SNPs (A: $G_{\alpha\eta}$, C: $G_{\alpha\beta}$) and the bias-corrected genetic correlation estimates (B: $G^A_{\alpha\eta}$, D: $G^A_{\alpha\beta}$). We set $h^2_\alpha = h^2_\beta = h^2_\eta = 1$, $n_1 = n_2 = p = 10,000$, and $n_3 = m = 2000$.

Theorem 3.2. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$ as $min(n_1, n_2, n_3, p) \to \infty$, and let $p^2 = c \cdot (n_1 n_2 n_3)^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then we have*

$$G_{\alpha\beta} = \varphi_{\alpha\beta} + \left(\sqrt{\frac{n_1}{n_1 + p/h^2_\alpha} \cdot \frac{n_2}{n_2 + p/h^2_\beta}} - 1 \right) \cdot \varphi_{\alpha\beta} + o_p(1).$$

If $a \in [1, \infty]$, then we have

$$G_{\alpha\beta} \cdot \frac{n_3(n_1 + p)(n_2 + p)}{p^2} = O_p(1).$$

Remark 3.3. *If $a \in (0, 1)$, $G_{\alpha\beta}$ is an unbiased estimator of $\varphi_{\alpha\beta}$ for $p = o\{\min(n_1, n_2)\}$. When $p = O(n_1) = O(n_2)$, $\sqrt{n_1/(n_1 + p/h^2_\alpha) \cdot n_2/(n_2 + p/h^2_\beta)}$ is smaller than 1, and thus $G_{\alpha\beta}$ is biased towards zero. Further if $\min(n_1, n_2) = o(p)$, $G_{\alpha\beta}$ is asymptotically zero. Therefore, to have nonzero asymptotic limit, both of the two sets of summary statistics need to be trained*

from large-scale GWAS. Giving that n_1, n_2, and p are proportional, the scale of $Var(G_{\alpha\beta})$ is

$$Var(G_{\alpha\beta}) = O_p\Big\{\frac{1}{n_3} + \frac{m_{\alpha\beta}}{m_\alpha m_\beta}\Big\} = O_p\Big\{\max(\frac{1}{n_3}, \frac{1}{m_{\alpha\beta}})\Big\},$$

and a consistent estimator of $\varphi_{\alpha\beta}$ is given by

$$G^A_{\alpha\beta} = G_{\alpha\beta} \cdot \sqrt{\frac{(n_1 + p/h^2_\alpha) \cdot (n_2 + p/h^2_\beta)}{n_1 n_2}} = \varphi_{\alpha\beta} + o_p(1).$$

Now we propose and study a novel estimator of $\varphi_{\alpha\beta}$ that can be directly constructed by using two sets of summary statistics $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$. Let

$$\widehat{\varphi}_{\alpha\beta} = \frac{\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\alpha}}\| \cdot \|\widehat{\boldsymbol{\beta}}\|} = \frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{\big\|\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\big\| \cdot \big\|\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\big\|},$$

we have the following asymptotic properties.

Theorem 3.3. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$ as $min(n_1, n_2, p) \to \infty$, and let $p = c \cdot (n_1 n_2)^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then we have*

$$\widehat{\varphi}_{\alpha\beta} = \varphi_{\alpha\beta} + \left(\sqrt{\frac{n_1}{n_1 + p/h^2_\alpha} \cdot \frac{n_2}{n_2 + p/h^2_\beta}} - 1\right) \cdot \varphi_{\alpha\beta} + o_p(1).$$

If $a \in [1, \infty]$, then we have $\widehat{\varphi}_{\alpha\beta} \cdot \{(n_1 + p)(n_2 + p)\}/p = O_p(1)$.

The $\widehat{\varphi}_{\alpha\beta}$ is interesting in its own right because it quantifies the potential bias of the inner product of marginal screening estimates in high-dimensions. When n_1, n_2, and p are proportional, $\mathrm{Var}(\widehat{\varphi}_{\alpha\beta}) = O_p\big\{\max(n_1^{-1}, m_{\alpha\beta}^{-1})\big\}$ and a consistent estimator of $\varphi_{\alpha\beta}$ is given by

$$\widehat{\varphi}^A_{\alpha\beta} = \widehat{\varphi}_{\alpha\beta} \cdot \sqrt{\frac{(n_1 + p/h^2_\alpha) \cdot (n_2 + p/h^2_\beta)}{n_1 n_2}} = \varphi_{\alpha\beta} + o_p(1).$$

Since $\widehat{\varphi}_{\alpha\beta}$ and $G_{\alpha\beta}$ have similar asymptotic properties, in what follows we will focus on $G_{\alpha\beta}$

and the general conclusions of $G_{\alpha\beta}$ remain the same for $\widehat{\varphi}_{\alpha\beta}$.

3.2 SNP screening

As shown in Theorems 3.1 and 3.2, in addition to heritability, the asymptotic limit of $G_{\alpha\eta}$ or $G_{\alpha\beta}$ is largely affected by n/p. These results intuitively suggest to select a subset of p SNPs to construct cross-trait PRS. The common approach in practice is to screen the SNPs according to their GWAS p-values. We investigate this strategy in this section.

For a given threshold $c_\alpha > 0$, let $q_\alpha = p \cdot \pi_\alpha = q_{\alpha 1} + q_{\alpha 2}$ ($\pi_\alpha \in (0, 1]$) be the number of top-ranked SNPs selected for $\boldsymbol{y}_\alpha$, among which there are $q_{\alpha 1}$ true causal SNPs and the remaining $q_{\alpha 2}$ are null SNPs, and we let $q_{\alpha\eta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$, and thus $q_{\alpha 1} \geq q_{\alpha\eta}$. The SNP data are defined accordingly. We write $\boldsymbol{X}_{(1)} = [\boldsymbol{X}_{(11)}, \boldsymbol{X}_{(12)}]$, $\boldsymbol{X}_{(2)} = [\boldsymbol{X}_{(21)}, \boldsymbol{X}_{(22)}]$, $\boldsymbol{W}_{(1,\alpha)} = [\boldsymbol{W}_{(11,\alpha)}, \boldsymbol{W}_{(12,\alpha)}]$, and $\boldsymbol{W}_{(2,\alpha)} = [\boldsymbol{W}_{(21,\alpha)}, \boldsymbol{W}_{(22,\alpha)}]$. Here $\boldsymbol{X}_{(11)}$ and $\boldsymbol{W}_{(11,\alpha)}$ are the selected $q_{\alpha 1}$ causal SNPs of $\boldsymbol{y}_\alpha$, and similarly, $\boldsymbol{X}_{(21)}$ and $\boldsymbol{W}_{(21,\alpha)}$ are the selected $q_{\alpha 2}$ null SNPs of $\boldsymbol{y}_\alpha$. In addition, we let $\widehat{\boldsymbol{\alpha}}_{(1)}^T = [\widehat{\boldsymbol{\alpha}}_{(11)}^T, \widehat{\boldsymbol{\alpha}}_{(12)}^T]$, and $\widehat{\boldsymbol{\alpha}}_{(2)}^T = [\widehat{\boldsymbol{\alpha}}_{(21)}^T, \widehat{\boldsymbol{\alpha}}_{(22)}^T]$, where $\widehat{\boldsymbol{\alpha}}_{(11)}$ corresponds to the selected causal SNPs of $\boldsymbol{y}_\alpha$, and $\widehat{\boldsymbol{\alpha}}_{(21)}$ corresponds to the selected null ones. Then we have

$$G_{T\alpha\eta} = \frac{\big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \big(\boldsymbol{W}_{(11,\alpha)}\widehat{\boldsymbol{\alpha}}_{(11)} + \boldsymbol{W}_{(21,\alpha)}\widehat{\boldsymbol{\alpha}}_{(21)}\big)}{\big\|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big\| \cdot \big\|\boldsymbol{W}_{(11,\alpha)}\widehat{\boldsymbol{\alpha}}_{(11)} + \boldsymbol{W}_{(21,\alpha)}\widehat{\boldsymbol{\alpha}}_{(21)}\big\|} = \frac{C_{T\alpha\eta}}{V_\eta \cdot V_{T\alpha}}$$

where $V_\eta = \big\|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big\|$, $V_{T\alpha} = \big\|\boldsymbol{W}_{(11,\alpha)}\boldsymbol{X}_{(11)}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) + \boldsymbol{W}_{(21,\alpha)}\boldsymbol{X}_{(21)}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\big\|$, and

$$C_{T\alpha\eta} = \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \boldsymbol{W}_{(11,\alpha)}\boldsymbol{X}_{(11)}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) + \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \boldsymbol{W}_{(21,\alpha)}\boldsymbol{X}_{(21)}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big).$$

Corollary 3.2. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose that min*$(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$ *and min*$(q_{\alpha\eta}, q_{\alpha 1}, q_{\alpha 2}) \to \infty$ *as min*$(n_1, n_3, p) \to \infty$*, further if* $\big\{m_{\alpha\eta}^2(q_{\alpha 1} +$

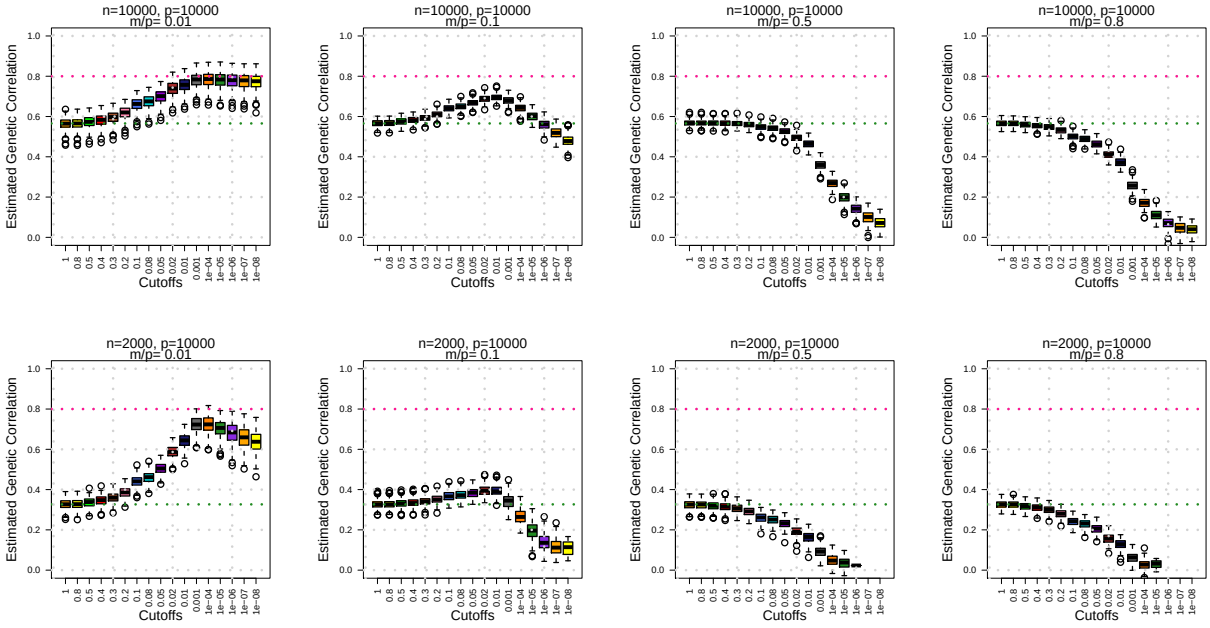

Figure 3.3: Raw genetic correlation $G_{T\alpha\eta}$ estimated by cross-trait PRS with selected SNPs under different sparsity m/p and sample size n. We set $h^2_\alpha = h^2_\eta = 1$, $\varphi_{\alpha\eta} = 0.8$, $p = 10,000$, and training sample size $n = 10,000$ (upper panels) or 2000 (lower panels).

$q_{\alpha 2})\big\}/(q^2_{\alpha\eta} n_1 n_3) \to 0$, *then we have*

$$G_{T\alpha\eta} = \varphi_{\alpha\eta} + \left(\sqrt{\frac{n_1 m_\alpha}{n_1 q_{\alpha 1} + m_\alpha q_\alpha / h^2_\alpha}} \cdot \frac{q_{\alpha\eta}}{m_{\alpha\eta}} \cdot h_\eta - 1 \right) \cdot \varphi_{\alpha\eta} + o_p(1).$$

Corollary 3.2 shows the trade-off of SNP screening. Given n_1, m_α, $m_{\alpha\eta}$, h_α, and h_η, the potential bias of $G_{T\alpha\eta}$ is also affected by q_α, $q_{\alpha 1}$ and $q_{\alpha\eta}$. As more SNPs are selected, the numerator of $\sqrt{(n_1 m_\alpha)/(n_1 q_{\alpha 1} + m_\alpha q_\alpha / h^2_\alpha)} \cdot (q_{\alpha\eta}/m_{\alpha\eta})$ increases with $q_{\alpha\eta}$, while the denominator increases with $\sqrt{q_\alpha}$ (and $\sqrt{q_{\alpha 1}}$). Therefore, whether or not SNP screening can improve the estimation is largely affected by the quality of the selected SNPs, which is highly related to the m_α/n_1 ratio. In the optimistic case where $q_{\alpha\eta} = m_{\alpha\eta}$ and $q_\alpha = q_{\alpha 1} = m_\alpha$, $G_{T\alpha\eta}$ becomes

$$\sqrt{\frac{n_1}{n_1 + m_\alpha / h^2_\alpha}} \cdot h_\eta \cdot \varphi_{\alpha\eta},$$

which is the theoretical upper limit. We note that this optimistic upper limit can still be biased towards zero. An opposite case is that the GWAS summary statistics of causal and null SNPs are totally mixed up, which may occur when m_α/n_1 is large (i.e., sample size is relatively small or trait is highly polygenic). Therefore, we have $q_{\alpha 1}/q_\alpha \approx m_\alpha/p$. Suppose also $q_{\alpha\eta}/q_{\alpha 1} \approx m_{\alpha\eta}/m_\alpha$, we have

$$G_{T\alpha\eta} \approx \sqrt{\frac{n_1}{n_1 p + p^2/h_\alpha^2} \cdot q_\alpha} \cdot h_\eta \cdot \varphi_{\alpha\eta},$$

which increases with q_α. As $q_\alpha = p$, $G_{T\alpha\eta}$ reaches its upper bound

$$\sqrt{\frac{n_1}{n_1 + p/h_\alpha^2}} \cdot h_\eta \cdot \varphi_{\alpha\eta}.$$

That is, $G_{T\alpha\eta}$ achieves the best performance when the cross-trait PRS is constructed without SNP screening. For example, in the left two panels of Figure 3.3, we set $m_\alpha/n_1 = 0.01$ (upper) and 0.05 (lower) to reflect the sparse signal cases, in which causal and null SNPs can be easily separated by SNP screening. Thus, SNP screening can reduce the bias of $G_{\alpha\eta}$ when signals are sparse. However, as the number of causal SNPs increase (from left to right in Figure 3.3), it becomes much hard to separate causal and null SNPs by their GWAS p-values. Therefore, SNP screening will enlarge the bias.

In conclusion, when causal and null SNPs can be easily separated by GWAS, the top-ranked SNPs are more likely to be causal ones, that is, SNP screening helps. However, for highly polygenic complex traits whose m_α/n_1 is large, SNP screening may result in larger bias. Moreover, since different underlying m_α/n_1 ratio will result in different patterns as shown in Figure 3.3, the observed pattern can be used to infer the m_α/n_1 ratio (i.e., the degree of polygenicity) and minimize the potential bias in estimation. We display this strategy using a real data example in Section 3.5. The $G_{\alpha\beta}$ has similar properties when performing SNP screening, whose results can be found in Appendix A.

3.3 Overlapping samples

In real data applications, different GWAS may share a subset of participants. It is often inconvenient to recalculate the GWAS summary statistics after removing the overlapping samples. In this section, we examine the effect of overlapping samples on the bias of cross-trait PRS, which provides more insights into the bias phenomenon of cross-trait PRS. Particularly, we focus on one case which is common in practice: n_s overlapping samples between discovery GWAS and Target testing data for $\varphi_{\alpha\eta}$ estimation. We add n_s overlapping samples into Discovery GWAS-I and Target testing GWAS, resulting in the following two new datasets:

- Dataset IV: $(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{y}_\alpha)$, with $\boldsymbol{X} \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{S} \in \mathbb{R}^{n_s \times p}$, and $\boldsymbol{y}_\alpha^T = (\boldsymbol{y}_{\alpha_X}^T, \boldsymbol{y}_{\alpha_S}^T) \in \mathbb{R}^{(n_1+n_s)\times 1}$.
- Dataset V: $(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{y}_\eta)$, with $\boldsymbol{W} \in \mathbb{R}^{n_3 \times p}$, $\boldsymbol{S} \in \mathbb{R}^{n_s \times p}$, and $\boldsymbol{y}_\eta^T = (\boldsymbol{y}_{\eta_W}^T, \boldsymbol{y}_{\eta_S}^T) \in \mathbb{R}^{(n_3+n_s)\times 1}$.

Mimicking h^2, we define $h_{\alpha\eta} \in (0, 1]$ as the proportion of phenotypic correlation that can be explained by the correlation of their genetic components as $h_{\alpha\eta} = (m_{\alpha\eta}/p)\sigma_{\alpha\eta}/\{(m_{\alpha\eta}/p)\sigma_{\alpha\eta} + \sigma_{\epsilon_\alpha\epsilon_\eta}\}$. On the overlapping samples, we allow nonzero correlation between random errors to capture the non-genetic contribution to phenotypic correlation. We introduce an additional condition on random errors.

Condition 3.4. *On n_s overlapping samples, ϵ_{α_j} and ϵ_{η_j} are independent random variables satisfying*

$$\begin{pmatrix} \epsilon_{\alpha_j} \\ \epsilon_{\eta_j} \end{pmatrix} \sim F\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\epsilon_\alpha}^2 & \sigma_{\epsilon_\alpha\epsilon_\eta} \\ \sigma_{\epsilon_\alpha\epsilon_\eta} & \sigma_{\epsilon_\eta}^2 \end{pmatrix} \right]$$

for $j = 1, ..., n_s$, where $\sigma_{\epsilon_\alpha\epsilon_\eta} = \rho_{\epsilon_\alpha\epsilon_\eta} \cdot \sigma_{\epsilon_\alpha}\sigma_{\epsilon_\eta}$.

Theorem 3.4. *Under polygenic model (3.4) and Conditions 3.1 - 3.4, suppose min($m_{\alpha\eta}$, m_α, m_η) $\to \infty$ as min$\{(n_1 + n_s), (n_3 + n_s), p\} \to \infty$, and let $p = c \cdot \{(n_1 + n_s)(n_3 + n_s)\}^a$ for*

some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then $G_{S\alpha\eta}$ can be written as

$$\frac{\left[1 + n_s p/\{(n_1 + n_s)(n_3 + n_s) \cdot h_{\alpha\eta}\}\right] \cdot \left[h_\eta \cdot \varphi_{\alpha\eta} \cdot \{1 + o_p(1)\}\right]}{\left[1 + p/\{(n_1 + n_s) \cdot h_\alpha^2\} + 2n_s p/\{(n_1 + n_s)(n_3 + n_s)\} + n_s p^2/\{(n_1 + n_s)^2(n_3 + n_s) \cdot h_\alpha^2\}\right]^{1/2}}.$$

If $a \in [1, \infty]$, then we have $G_{S\alpha\eta} = o_p(1)$.

Remark 3.4. *Theorem 3.4 shows the effect of n_s overlapping samples on the estimation of $\varphi_{\alpha\eta}$. Both sample sizes $(n_1 + n_s)$ and $(n_3 + n_s)$ are involved in the limit. An interesting special case is when the two GWAS are fully overlapped, then we have*

$$G_{S\alpha\eta} = \frac{n_s + p/h_{\alpha\eta}}{\left\{n_s^2 + 2n_s p + p(p + n_s)/h_\alpha^2\right\}^{1/2}} \cdot h_\eta \cdot \varphi_{\alpha\eta} + o_p(1).$$

In the optimal situation where $h_\alpha^2 = h_\eta^2 = h_{\alpha\eta} = 1$, we have

$$G_{S\alpha\eta} = \left(1 + \frac{1}{p/n_s + n_s/p + 2}\right)^{-1/2} \cdot \varphi_{\alpha\eta} + o_p(1).$$

Therefore, $G_{S\alpha\eta}$ is asymptotically biased unless either $p = o(n_s)$ or $n_s = o(p)$ holds, neither of which is the case in modern GWAS. As n_s and p are more comparable, the asymptotic bias in $G_{S\alpha\eta}$ increases and the largest bias occurs as $p = n_s \to \infty$.

Note that it is not recommended to estimate the genetic correlation between two traits with (fully) overlapping samples due to concerns such as confounding and overfitting (Pasaniuc and Price, 2017; Dudbridge, 2013). In our analysis, such concern is quantified by the value of $h_{\alpha\eta}$. That is, when non-genetic correlation exists in error terms, we have $h_{\alpha\eta} < 1$, and the estimation of genetic correlation is inflated. However, on the other hand, our results show that even in an optimal overlapping setting with $h_\alpha^2 = h_\eta^2 = h_{\alpha\eta} = 1$, the cross-trait PRS estimator based on GWAS summary statistics can be biased towards zero.

In Appendix A, we further investigate several other specific overlapping cases, which can be useful for quantifying potential bias and perform correction in real data. In summary, these analyses reveal that the bias in cross-trait PRS estimator may result from the following

facts: i) summary statistics are generated from independent GWAS, where the induced bias is largely determined by the n/p ratio; ii) phenotypes are not fully heritable, i.e., heritability is less than one; and iii) non-genetic correlation exists in the random errors of overlapping samples. This may happen, for example, when confounding effects are not fully adjusted. The first two facts may bias the genetic correlation estimator towards zero, while the last fact may inflate the estimated genetic correlation.

3.4 Numerical experiments

3.4.1 Cross-trait PRS with all SNPs

To illustrate the finite sample performance of our theoretical results, we simulate $10,000$ uncorrelated SNPs. The MAF of each SNP, f, is independently generated from Uniform $[0.05, 0.45]$ based on which the SNP genotypes are independently sampled from $\{0, 1, 2\}$ with probabilities $\{(1-f)^2, 2f(1-f), f^2\}$, respectively. The SNPs are then standardized to satisfy Condition 3.1. We set the same 2000 causal SNPs on each trait and the nonzero genetic effects are generated from Normal distribution according to Condition 3.3 with $\sigma_\alpha = \sigma_\eta = \sigma_\beta = 1$. We set all heritability to one and vary $\sigma_{\alpha\eta}$ and $\sigma_{\alpha\beta}$ (and thus asymptotically $\varphi_{\alpha\eta}$ and $\varphi_{\alpha\beta}$) from 0.1 to 0.9. Model (3.4) is used to generate continuous phenotypes. We generate $10,000$ samples in training dataset and 2000 samples in testing dataset. A total of 200 replicates was conducted. Cross-trait PRS is built with all SNPs. We calculate the raw estimators $G_{\alpha\eta}$ and $G_{\alpha\beta}$ studied in Theorems 3.1 - 3.2, and the corresponding bias-corrected estimators $G^A_{\alpha\eta}$ and $G^A_{\alpha\beta}$. The performance of $G_{\alpha\eta}$ and $G_{\alpha\beta}$ is displayed in the panels A and C of Figure 3.2. It is clear that these raw estimates are biased towards zero. For example, when $\sigma_{\alpha\eta} = \sigma_{\alpha\beta} = 0.9$, $G_{\alpha\eta}$ is around 0.6 while $G_{\alpha\beta}$ is less than 0.45. The performance of $G^A_{\alpha\eta}$ and $G^A_{\alpha\beta}$ is displayed in the panels B and D of Figure 3.2, which indicates that the two bias-corrected estimators perform well and are close to the true value of $\sigma_{\alpha\eta}$ and $\sigma_{\alpha\beta}$, respectively.

To verify that our results are independent of the signal sparsity, we set $m_\alpha = m_\beta = m_\eta = p \cdot a_\alpha$ and vary the sparsity $a_\alpha = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.6, 0.7$ and 0.8 to generate sparse

and dense signals. Next, we fix $a_\alpha = 0.2$ and set $m_\beta = m_\eta = k \cdot m_\alpha$ to allow phenotypes to have different number of causal SNPs, where $k = 0.3, 0.4, 0.5, 0.8, 1, 1.25, 2, 2.5$ and 3.3. We set all heritability to one and let $\sigma_{\alpha\eta} = \sigma_{\alpha\beta} = 0.5$. Sample size of training and testing datasets is set to either 2000 or $10,000$. The performance of $G_{\alpha\eta}$ is displayed in the upper panels of Figure 3.4. The bias of $G_{\alpha\eta}$ is independent of the sparsity a_α of a trait or the ratio of sparsity k between two traits, which verifies our results of Theorem 3.1. The bottom panels of Figure 3.4 display the performance of $G^A_{\alpha\eta}$. It is clear that $G^A_{\alpha\eta}$ is unbiased regardless of a_α and k. The Figure 3.5 shows a similar pattern in $G^A_{\alpha\eta}$ as heritability $h^2_\alpha = h^2_\eta = 0.5$. The performance of $G_{\alpha\beta}$ and $G^A_{\alpha\beta}$ is displayed in Figure 3.6 and supports our results in Theorem 3.2. Finally, we illustrate the performance of $\widehat{\varphi}_{\alpha\beta}$ and $\widehat{\varphi}^A_{\alpha\beta}$ in Figure 3.7, verifying our results in Theorem 3.3 and the unbiasedness of $\widehat{\varphi}^A_{\alpha\beta}$.

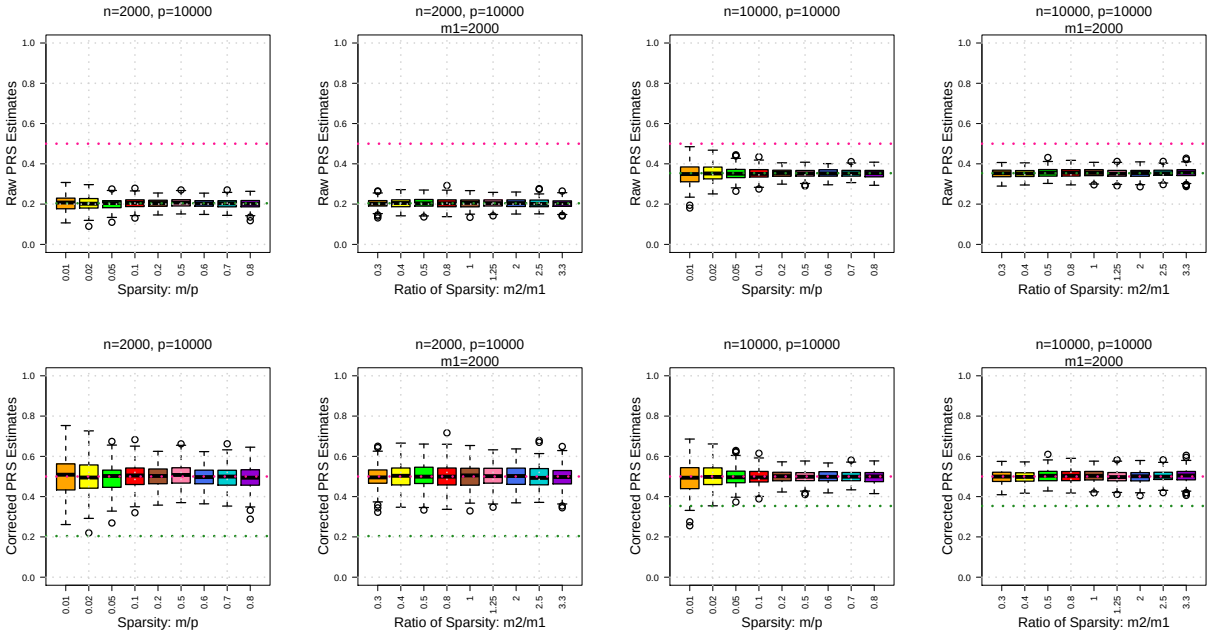

Figure 3.4: Raw genetic correlations estimated by cross-trait PRS with all SNPs ($G_{\alpha\eta}$, upper panels) and corrected ones based on our formulas ($G^A_{\alpha\eta}$, bottom panels). We set $h^2_\alpha = h^2_\eta = 1$, $\varphi_{\alpha\eta} = 0.5$, $p = 10,000$, and vary m_α, m_η and n.

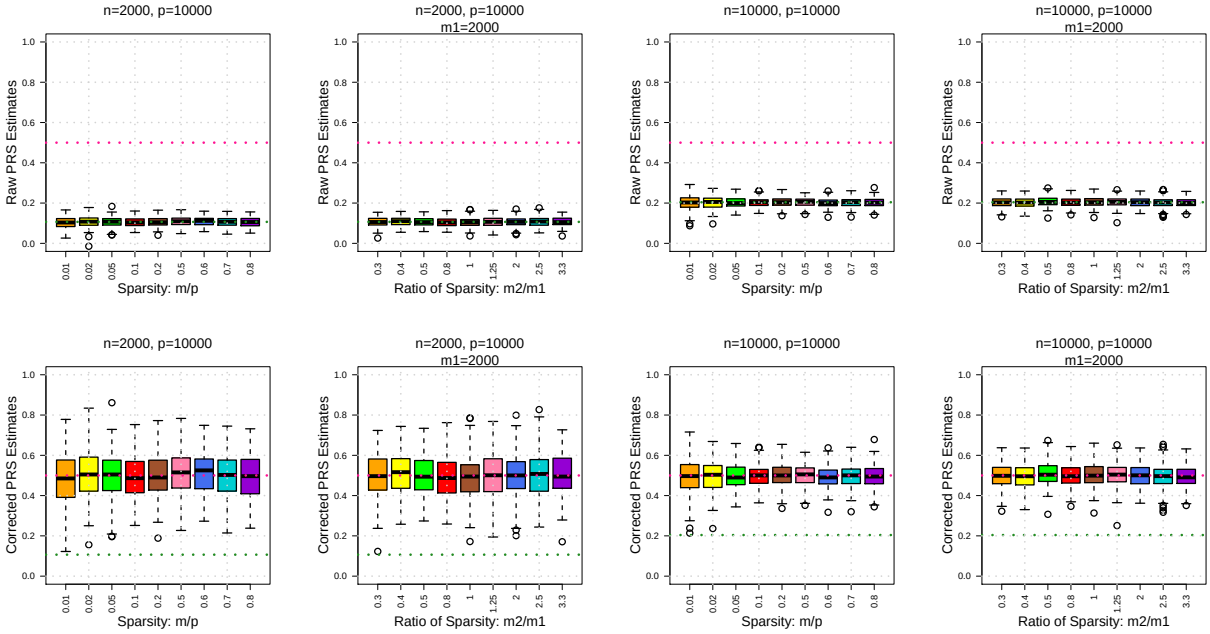

Figure 3.5: Raw genetic correlations estimated by cross-trait PRS with all SNPs ($G_{\alpha\eta}$, upper panels) and corrected ones based on our formulas ($G^A_{\alpha\eta}$, bottom panels). We set $h^2_\alpha = h^2_\eta = 0.5$, $\varphi_{\alpha\eta} = 0.5$, $p = 10,000$, and vary m_α, m_η and n.

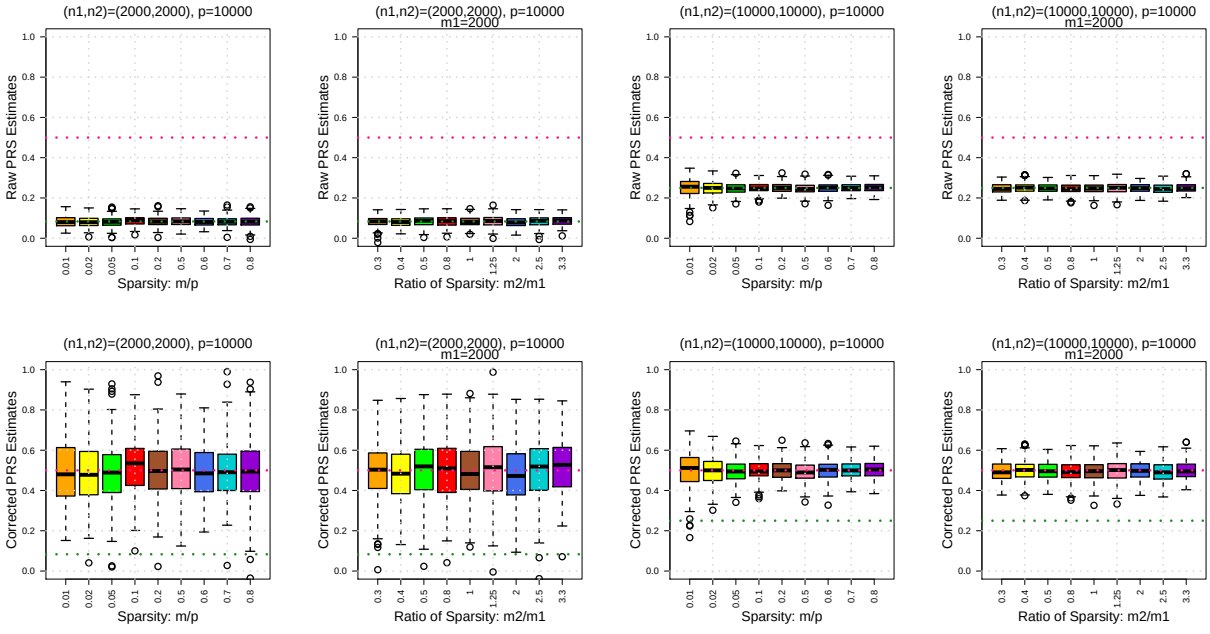

Figure 3.6: Raw genetic correlations estimated by cross-trait PRS with all SNPs ($G_{\alpha\beta}$, upper panels) and corrected ones based on our formulas ($G^A_{\alpha\beta}$, bottom panels). We set $h^2_\alpha = h^2_\beta = 1$, $\varphi_{\alpha\beta} = 0.5$, $p = 10,000$, and vary m_α, m_β and n.

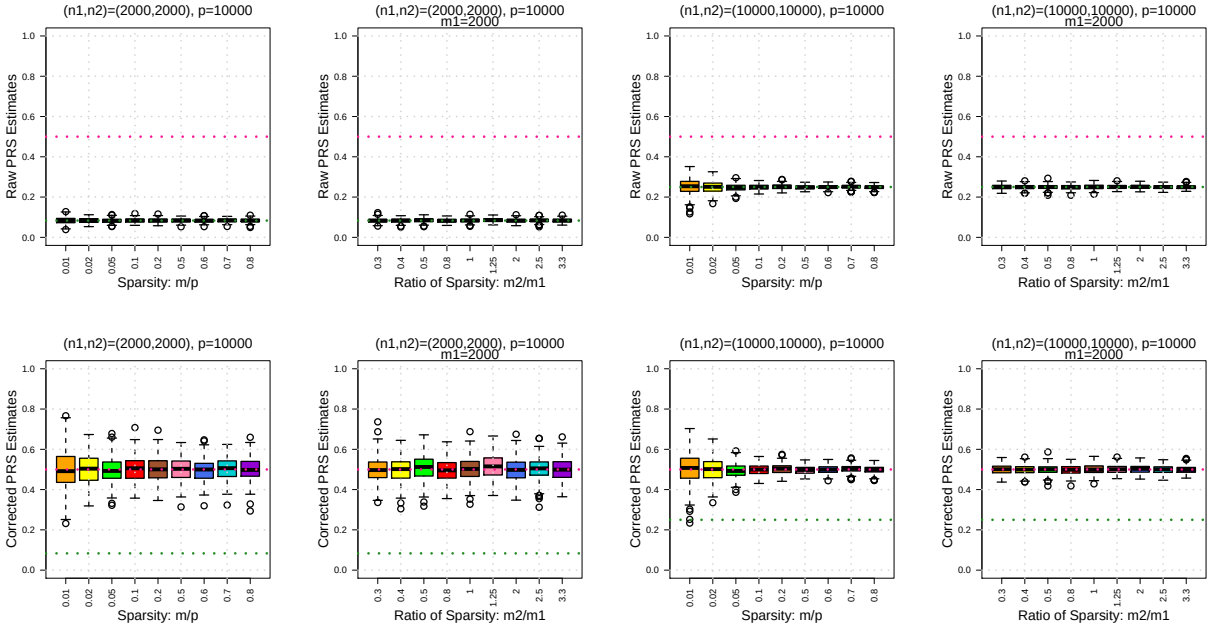

Figure 3.7: Raw genetic correlations estimated by cross-trait PRS directly with all SNPs ($\widehat{\varphi}_{\alpha\beta}$, upper panels) and corrected ones based on our formulas ($\widehat{\varphi}^{A}_{\alpha\beta}$, bottom panels). We set $h^2_\alpha = h^2_\beta = 1$, $\varphi_{\alpha\beta} = 0.5$, $p = 10,000$, and vary m_α, m_β and n.

3.4.2 SNP screening and overlapping samples

Instead of using all the $10,000$ SNPs, we construct cross-trait PRS with the top-ranked SNPs whose GWAS p-values pass a pre-specified threshold. We consider a series of thresholds $\{1, 0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.05, 0.02, 0.01, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$ and generate a series of $G_{T\alpha\eta}$ accordingly. We set heritability to one and $\varphi_{\alpha\eta} = 0.8$. Four levels of sparsity $m_\alpha/p = m_\eta/p = 0.01, 0.1, 0.5$ and 0.8 are examined. Figure 3.3 displays the performance of $G_{T\alpha\eta}$ across a series of thresholds. As expected, the pattern of $G_{T\alpha\eta}$ varies dramatically with the sparsity. When signals are sparse, SNP screening helps and $G_{T\alpha\eta}$ performs better than $G_{\alpha\eta}$. However, when signals are dense, the performance of $G_{T\alpha\eta}$ drops as the threshold decreases. $G_{T\alpha\eta}$ has the best performance as all SNPs are selected, i.e., the same as $G_{\alpha\eta}$, which confirms our results of $G_{T\alpha\eta}$ in Corollary 3.2. In addition, we examine our analyses of overlapping samples. For $G_{S\alpha\eta}$ and $G_{S\alpha\beta}$, half of the $10,000$ samples are set to be overlapping. Other settings remain the same as those of Figure 3.2. The performance

of $G_{S\alpha\eta}$, $G_{S\alpha\beta}$, $G^A_{S\alpha\eta}$ and $G^A_{S\alpha\beta}$ is displayed in Figure 3.8, which fully supports the results in Theorems 3.4 and Proposition S4.

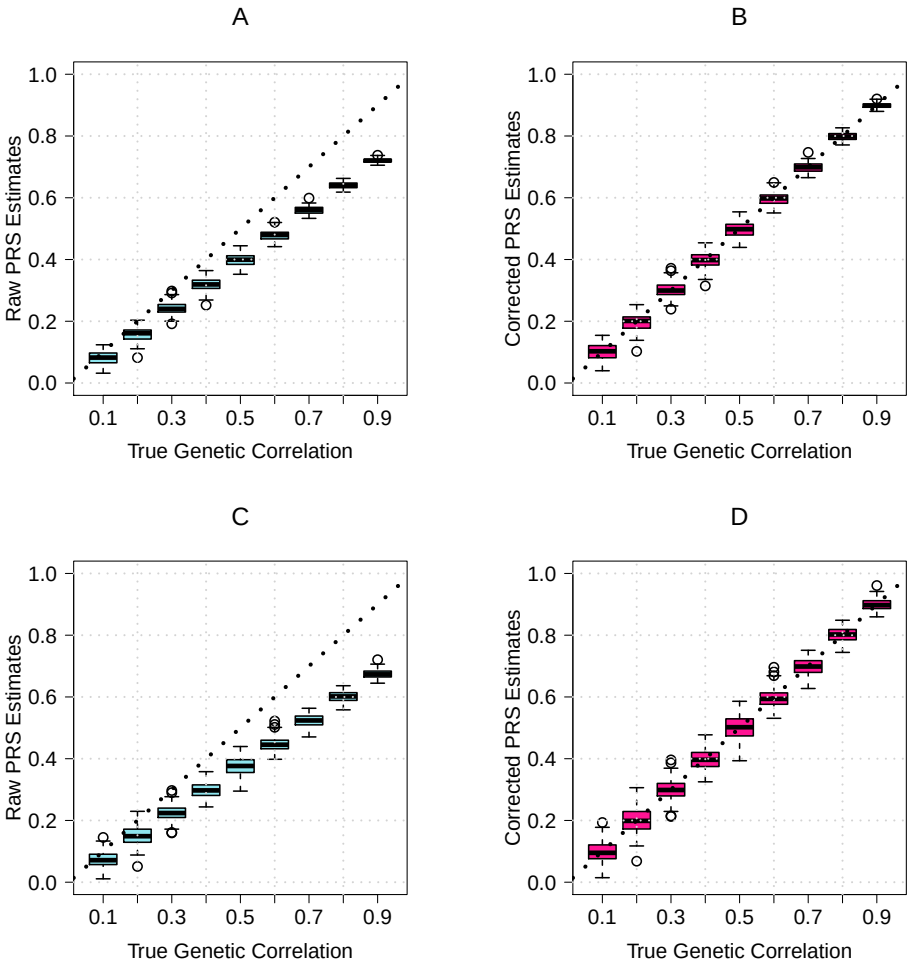

Figure 3.8: Raw genetic correlations estimated by cross-trait PRS with all SNPs (left panels, A: $G_{S\alpha\eta}$, C: $G_{S\alpha\beta}$) and bias-corrected genetic correlation estimates (right panels, B: $G^A_{S\alpha\eta}$, D: $G^A_{S\alpha\beta}$). We set $h^2_\alpha = h^2_\beta = h^2_\eta = 1$, $n_1 = n_s = n_2 = n_3 = 5000$ (half samples overlap), $p = 10,000$, and $m = 2000$.

3.5 Real data analysis

Human brain structural changes are known to be associated with cognitive and mental health traits (Caldiroli et al., 2018; Vreeker et al., 2017; Davies et al., 2016). It is an active research area to understand the shared genetic influences among these brain-related complex traits (Wei et al., 2019; Jansen et al., 2019). Volumes of brain region of interest (ROI) (refer

to as ROI volumes) are heritable measures of brain structural variation and can be quantified by brain magnetic resonance imaging (MRI). In this section, we assess the genetic correlation between ROI volumes and reaction time, which is a heritable measure of general cognitive functions (Davies et al., 2018). We focus on the volume measures from seven important brain ROIs, including thalamus proper, caudate, putamen, pallidum, hippocampus, accumbens area, and the total brain volume (TBV). These ROIs are frequently studied in imaging genetics (Hibar et al., 2015), and common SNPs are reported to be able to account for about 50% phenotypic variation in these traits (Biton et al., 2019).

As a positive control, we first estimate the genetic correlation between the TBV phenotype measured in the Pediatric Imaging, Neurocognition, and Genetics (PING) study (Jernigan et al., 2016) and the same trait measured in the United Kingdom Biobank (UKB) study (Sudlow et al., 2015). The TBV phenotype in the two studies is generated using consistent standard procedures via advanced normalization tools (ANTs, Avants et al. (2011)), and thus the underlying genetic correlation is expected to be close to one. A full description of the PING study, ANTs processing, genotyping data quality controls (QCs) is documented in Appendix A. We generate the PRS on our PING samples ($n = 924$) by summarizing across all the LD-pruned candidate SNPs ($R^2 = 0.2$, window size 50), weighed by the published UKB GWAS effect sizes (Zhao et al. (2019), $n = 19,629$, `https://github.com/BIG-S2/GWAS`). Plink tool set (Purcell et al., 2007) is used to generate these scores. The association between TBV and the constructed PRS is estimated and tested in linear regression, adjusting for the effects of age and sex. The additional phenotypic variation that can be explained by the PRS (i.e., the partial R^2) is interpreted as an estimator of the squared genetic correlation. The partial R^2 is 1.82% (p-value$=1.92 \times 10^{-6}$) in this positive control analysis, which is much smaller than one. This example illustrates that the genetic correlation estimated by the PRS can be heavily biased toward zero, but the testing power of PRS can still be good, which supports our theoretical results and matches many empirical observations.

Next, cross-trait PRS of reaction time is constructed on these PING samples using the

published GWAS summary statistics of reaction time from the largest study so far (Davies et al. (2018), $n = 282,014$, `https://www.ccace.ed.ac.uk/node/335`). The original GWAS has no overlapping samples with the PING study. We examine the partial R^2 using the same procedure as in the above positive control analysis. The results are summarized in Table 3.1. The mean proportion of variation that can be additionally explained by the cross-trait PRS is 1.31% across the seven ROIs. The largest partial R^2 2.80% is found in the thalamus region (p-value$=9.46 \times 10^{-9}$), which is known to play integrative roles in cognitive functions (Wolff and Vann, 2019). Evidence from imaging studies indicates that thalamus is phenotypically associated with reaction time (Brücke et al., 2013) and has predictive power to this cognitive trait (Nikulin et al., 2008). However, though the p-value reveals significant genetic relationship between thalamus and reaction time, the partial R^2 is small and may under-interpret the genetic similarity of the two traits. Thus, we correct the observed partial R^2 with our formula in Theorem 3.1. We use the heritability estimate of reaction time reported in Davies et al. (2018) ($h^2 = 0.25$), and the heritability estimates of ROI volumes reported in Biton et al. (2019). The number of independent variants p is estimated from the above positive control analysis of TBV, which is about $348,374$. After correction, the partial R^2 of TBV becomes 10.00%.

Table 3.1: Pairwise genetic correlation between the seven ROI volumes and reaction time estimated by cross-trait PRS.

ROI ID	p-value	Partial R^2	Heritability	Corrected partial R^2
thalamus proper	9.463×10^{-9}	2.792×10^{-2}	0.501	0.331
caudate	1.718×10^{-3}	1.033×10^{-2}	0.570	0.108
putamen	1.926×10^{-3}	9.249×10^{-3}	0.455	0.121
pallidum	2.810×10^{-2}	4.390×10^{-3}	0.340	0.077
hippocampus	3.053×10^{-6}	1.700×10^{-2}	0.312	0.324
accumbens area	2.937×10^{-4}	1.281×10^{-2}	0.300	0.254
total brain volume	5.117×10^{-4}	9.725×10^{-3}	0.578	0.100

We then apply correction with the number $348,374$ for the other six ROI volumes. The mean partial R^2 across the seven ROIs becomes 18.77%, which indicates a moderate level of

genetic correlation between reaction time and these ROI volumes in PING samples. These results are summarized in Table 3.1.

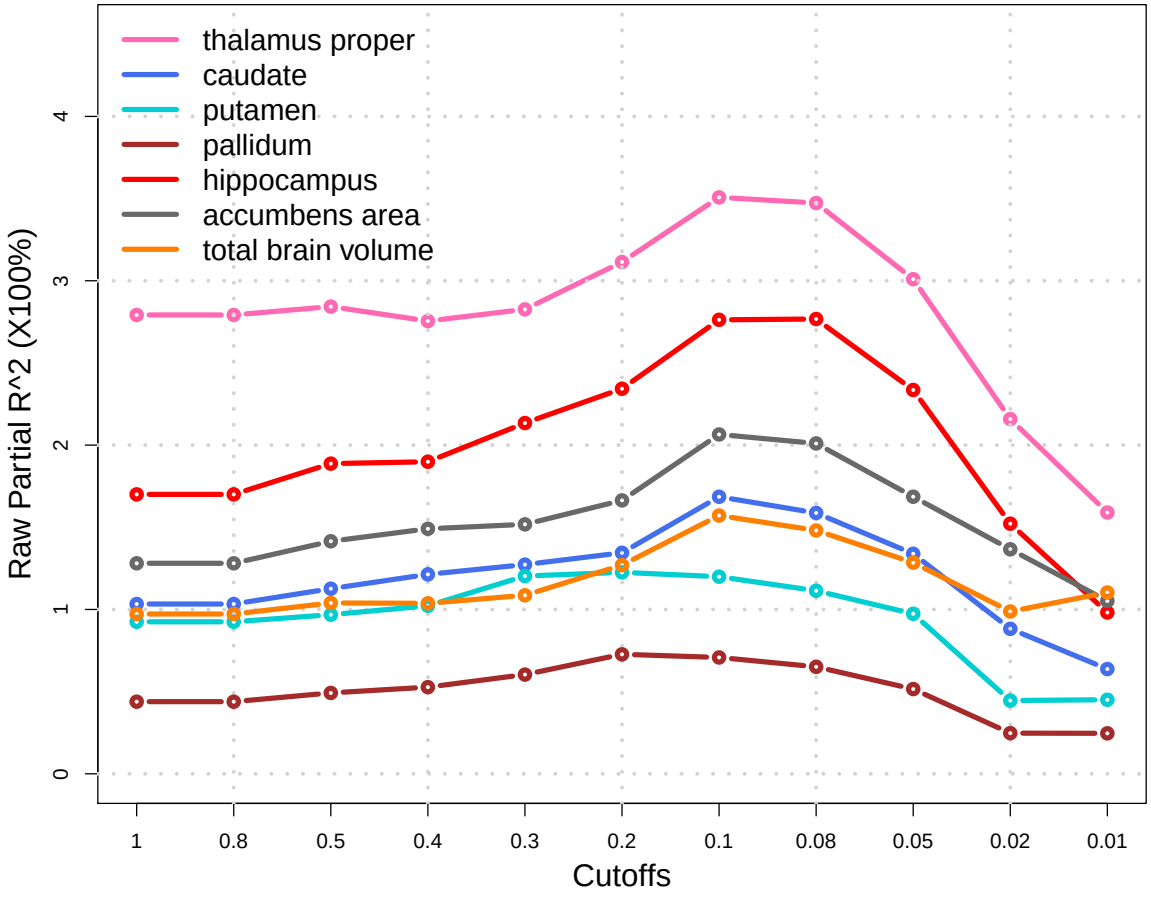

Figure 3.9: Raw partial R^2 of fitting reaction time PRS on seven regional brain volumes (listed in the figure) in the PING study given different GWAS p-value cutoffs.

To uncover the underlying m/n ratio and minimize the potential bias in the raw partial R^2, we apply SNP screening with multiple GWAS p-value cutoffs and present the trajectory of partial R^2 in Figure 3.9. The pattern of partial R^2 is similar across the seven ROI volumes, suggesting that these ROI volumes have similar genetic co-architecture with reaction time. The optimal GWAS p-value cutoff for genetic correlation estimation is around 0.1 in this analysis. The partial R^2 of thalamus moves up to 3.51% given this cutoff.

In summary, we examine the genetic correlation between reaction time and seven brain

ROI volumes in the PING study. Compared to the raw partial R^2, the corrected R^2 may better reflect the degree of genetic similarity between the two kinds of brain-related traits and suggests the potential prediction power of these brain imaging markers to cognitive functions. The trajectory in Figure 3.9 indicates that the genetic co-architecture of reaction time and ROI volumes has moderate to high level of polygenicity, which indicates that a large number of common genetic variants that simultaneously contribute to these traits. This result matches recent findings that brain-related traits can be highly polygenic and genetically related (Biton et al., 2019; Jansen et al., 2019; O'Connor et al., 2019; Zhao et al., 2019). Finally, we note that methods for genetic correlation estimation based on two sets of GWAS summary statistics, such as cross-trait LDSC, are known to require all the input summary statistics from large-scale discovery GWAS. Thus, their results can be noisy when the testing GWAS sample size is small (Ni et al., 2018), e.g., smaller than 5000 as mentioned in `https://github.com/bulik/ldsc/wiki/FAQ`.

3.6 Discussion

Understanding the genetic similarity among human complex traits is essential to model biological mechanisms, improve genetic risk prediction, and design personalized prevention/treatment. Cross-trait PRS (Purcell et al., 2009; Power et al., 2015) is one of the most popular methods for genetic correlation estimation with thousands of publications. This chapter empirically and theoretically studies the asymptotic properties of cross-trait PRS. Our analyses demystify the commonly observed small R^2 in real data applications, and help avoid over- or under-interpreting of research findings. The asymptotic behavior of cross-trait PRS and the performance of SNP screening are closely related to the ratios among (n, m, p). More importantly, the asymptotic bias is largely independent of the unknown genetic architecture if we use all SNPs in cross-trait PRS, which enables bias correction. As more discovery GWAS summary statistics from biobanks become publicly available (Watanabe et al., 2018), our bias-corrected estimators can be used to assess the underlying genetic correlation of many complex traits, especially when the in-house testing GWAS is relatively small. We also

discuss the SNP screening strategy and illustrate that the screening pattern can be used to assess the polygenicity and minimize the potential bias in real data applications. Influence of overlapping samples is also quantified in several practical cases. The training-testing design employed by cross-trait PRS may help avoid the inflation caused by non-genetic correlation, but results in systematic bias due to the restricted out-of-sample power of GWAS summary statistics in an independent testing dataset.

CHAPTER 4: GENETICS PREDICTION OF COMPLEX TRAITS

4.1 Preliminaries

In this section, we introduce the modeling framework, including the genetic architecture, assumptions on SNP data and genetic effects. We also introduce some useful RMT lemmas. We make use of the following notations frequently. $\mathrm{tr}(\boldsymbol{A})$ is the trace of matrix $\boldsymbol{A}$, $\mathrm{Diag}(\boldsymbol{A})$ is the diagonal of matrix $\boldsymbol{A}$, $\boldsymbol{A}^-$ is the inverse of matrix $\boldsymbol{A}$, $\boldsymbol{A}^T$ is the transpose of matrix $\boldsymbol{A}$, and $\boldsymbol{A}^+$ is the Moore-Penrose pseudoinverse of matrix $\boldsymbol{A}$. $\to$ donates the convergence of a series of real numbers, $\to_p$ represents the in probability convergence of a series of random variables, and $\to_{a.s.}$ is the almost surely convergence of a series of random variables. $\lambda_i(\boldsymbol{A})$ is the ith eigenvalue of matrix $\boldsymbol{A}$, $\mathrm{I}(\cdot)$ is the indicator function, and $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x} = \sum_{i=1}^p x_i^2$ is the squared l_2 norm of $p \times 1$ vector $\boldsymbol{x}$, and $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x}$ is the norm induced by $\boldsymbol{\Sigma}$. In addition, $o(1)$ and $O(1)$ define the small o and big O, $o_p(1)$ and $O_p(1)$ define the small o and big O in probability, and c, C are some generic constant numbers.

4.1.1 Modeling framework

Cross-trait prediction Consider two independent GWAS that are conducted for two traits with the same p SNPs (features):

- Training GWAS: $(\boldsymbol{X}, \boldsymbol{y})$, with $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] \in \mathbb{R}^{n\times p}$, $\boldsymbol{X}_{(1)} \in \mathbb{R}^{n\times m_\beta}$, and $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$.
- Testing GWAS: $(\boldsymbol{Z}, \boldsymbol{y}_z)$, with $\boldsymbol{Z} = [\boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)}] \in \mathbb{R}^{n_z\times p}$, $\boldsymbol{Z}_{(1)} \in \mathbb{R}^{n_z\times m_\eta}$, and $\boldsymbol{y}_z \in \mathbb{R}^{n_z\times 1}$.

Here $\boldsymbol{y}$ and $\boldsymbol{y}_z$ are two continuous phenotypes measured in two independent groups of individuals with sample sizes n and n_z, respectively. The $\boldsymbol{X}_{(1)}$ is an $n \times m_\beta$ matrix of the SNP data with nonzero effects, and $\boldsymbol{X}_{(2)}$ is an $n \times (p - m_\beta)$ matrix of the null SNPs, resulting

in an $n \times p$ matrix of all SNPs, donated by $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{m_\beta}, \boldsymbol{x}_{m_\beta+1}, \cdots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ is an $n \times 1$ vector of the SNP i, $i = 1, \cdots, p$. Similarly, the $\boldsymbol{Z}_{(1)}$ denotes the causal SNPs of $\boldsymbol{y}_z$ and $\boldsymbol{Z}_{(2)}$ donate the null SNPs. We allow $\boldsymbol{y}$ and $\boldsymbol{y}_z$ to be two different traits. That is, we consider a general cross-trait prediction problem, such as predicting cognitive ability by educational attainment (Lee et al., 2018), and treat same-trait prediction as a special case. Thus, m_β and m_η can be different numbers and $\boldsymbol{X}_{(1)}$ and $\boldsymbol{Z}_{(1)}$ correspond to two different sets of causal SNPs in general. The linear polygenic model assumes

$$\boldsymbol{y} = \boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}, \quad \text{and} \quad \boldsymbol{y}_z = \boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z, \tag{4.1}$$

where $\boldsymbol{\beta}_{(1)}^T = (\beta_1, \cdots, \beta_{m_\beta})^T$ and $\boldsymbol{\eta}_{(1)}^T = (\eta_1, \cdots, \eta_{m_\eta})^T$ are vectors of nonzero causal SNP effects, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_z$ represent independent random error vectors. We let $\boldsymbol{\beta}^T = \left(\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T\right)$ and $\boldsymbol{\eta}^T = \left(\boldsymbol{\eta}_{(1)}^T, \boldsymbol{\eta}_{(2)}^T\right)$, in which elements in $\boldsymbol{\beta}_{(2)}^T = (\beta_{m_\beta+1}, \cdots, \beta_p)^T$ and $\boldsymbol{\eta}_{(2)}^T = (\eta_{m_\eta+1}, \cdots, \eta_p)^T$ are all zeros. We model $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\eta}_{(1)}$ as random variables (Dobriban and Wager, 2018) and will introduce the detailed distribution assumptions in the following section. The overall genetic heritability of $\boldsymbol{y}$ and $\boldsymbol{y}_z$ are given by

$$h_\beta^2 = \frac{\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta})}{\mathrm{Var}(\boldsymbol{y})} = \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}}{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}} \quad \text{and} \quad h_\eta^2 = \frac{\mathrm{Var}(\boldsymbol{Z}\boldsymbol{\eta})}{\mathrm{Var}(\boldsymbol{y}_z)} = \frac{\boldsymbol{\eta}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}_z^T\boldsymbol{\epsilon}_z}. \tag{4.2}$$

We assume h_β^2 and $h_\eta^2 \in (0, 1]$.

Model assumptions and definitions Since m_β and m_η can be different and the causal SNPs of different traits may partially overlap, we let $m_{\beta\eta} \leq \min(m_\beta, m_\eta)$ be the number of overlapping causal SNPs of $\boldsymbol{y}$ and $\boldsymbol{y}_z$.

SNP data The assumptions on SNP data $\boldsymbol{X}$ and $\boldsymbol{Z}$ are summarized in Condition 4.1.

Condition 4.1. *1. SNP data satisfy $\boldsymbol{X} = \boldsymbol{X}_0\boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{Z} = \boldsymbol{Z}_0\boldsymbol{\Sigma}^{1/2}$, and entries of $\boldsymbol{X}_0$ and $\boldsymbol{Z}_0$ are real-value i.i.d. random variables with mean zero, variance one and a finite 12th*

order moment. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *population level deterministic positive definite matrix with* $0 < c \leq \lambda_{min}(\boldsymbol{\Sigma}) \leq \lambda_{max}(\boldsymbol{\Sigma}) \leq C$ *for all* n, p *and some constants* c, C*, where* $\lambda_{min}(\boldsymbol{\Sigma})$ *and* $\lambda_{max}(\boldsymbol{\Sigma})$ *are the smallest and largest eigenvalues of* $\boldsymbol{\Sigma}$*, respectively.* $\boldsymbol{\Sigma}^{1/2}$ *is any nonnegative square root of* $\boldsymbol{\Sigma}$*. For simplicity, we assume* $\boldsymbol{\Sigma}_{ii} = 1$*,* $i = 1, \cdots, p$*, or equivalently,* $\boldsymbol{X}$ *and* $\boldsymbol{Z}$ *have been column-standardized. In summary,* $\boldsymbol{\Sigma}$ *is assumed to be a correlation matrix with uniformly bounded eigenvalues.*

2. *As* $min(n, n_z, m_{\beta\eta}) \to \infty$*, we assume*

$$m_\beta/n \to \gamma, \quad m_\eta/n_z \to \gamma_z, \quad for \quad \gamma, \gamma_z \in (0, \infty);$$
$$p/n \to \omega, \quad p/n_z \to \omega_z, \quad for \quad \omega, \omega_z \in (0, \infty); \quad and$$
$$m_{\beta\eta}/\sqrt{m_\beta m_\eta} \to \kappa_{\beta\eta}, \quad for \quad \kappa_{\beta\eta} \in (0, 1].$$

3. *For the* $p \times p$ *population level correlation matrix* $\boldsymbol{\Sigma}$*, we define its empirical spectral distribution (ESD) as* $F_p^{\boldsymbol{\Sigma}}(x) = p^{-1} \cdot \sum_{i=1}^{p} I(\lambda_i(\boldsymbol{\Sigma}) \leq x)$*,* $x \in \mathbb{R}$*. As* $p \to \infty$*, let* $\{\boldsymbol{\Sigma}_p\}_{p>1}$ *be a sequence of matrices, we assume the sequence of corresponding ESDs* $\{F_p^{\boldsymbol{\Sigma}}(x)\}_{p>1}$ *converges weakly to a limit probability distribution* $H(x)$*,* $x \in \mathbb{R}$*, named the limiting spectral distribution (LSD) of* $\boldsymbol{\Sigma}$*.*

Condition 4.1 is frequently used in high-dimensional data analysis (Ledoit and Péché, 2011; Dobriban and Wager, 2018; Hastie et al., 2019). Condition 4.1 (a) assumes that $\boldsymbol{X}$ and $\boldsymbol{Z}$ have the same $\boldsymbol{\Sigma}$. In Section 4.4, we relax this assumption and extend our results to the case where the $\boldsymbol{X}$ and $\boldsymbol{Z}$ have different correlation structure. In modern GWAS, the number of the SNPs p in Condition 4.1 (b) is usually in millions (Tam et al., 2019), and the current training GWAS simple size n is often much smaller than p. The genetics community has striven to increase the GWAS sample size in recent years, and for some traits the sample size has been larger than one million (Lee et al., 2018; Nielsen et al., 2018). With these great efforts, ω is decreasing from a large number to one, or even can be less than one.

Genetic effects and random errors Let $F(0, V)$ represent a generic distribution with mean zero, (co)variance V, and finite fourth order moments. We introduce the following condition on genetic effects $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\eta}_{(1)}$ and random error vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_z$.

Condition 4.2. *We assume the distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are independent of $\boldsymbol{\Sigma}$. Moreover, β_i and η_j are independent random variables satisfying*

$$\beta_i \sim F(0, \sigma_\beta^2/p), \quad i = 1, ..., m_\beta; \qquad \eta_j \sim F(0, \sigma_\eta^2/p), \quad j = 1, ..., m_\eta.$$

The $m_{\beta\eta}$ overlapping nonzero effects (β_k, η_k)s of $(\boldsymbol{y}, \boldsymbol{y}_z)$ satisfy

$$\begin{pmatrix} \beta_k \\ \eta_k \end{pmatrix} \sim F\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, p^{-1} \cdot \begin{pmatrix} \sigma_\beta^2 & \sigma_{\beta\eta} \\ \sigma_{\beta\eta} & \sigma_\eta^2 \end{pmatrix} \right],$$

where $\sigma_{\beta\eta} = \rho_{\beta\eta} \cdot \sigma_\beta \sigma_\eta$. And ϵ_i and ϵ_{z_j} are independent random variables satisfying

$$\epsilon_i \sim F(0, \sigma_\epsilon^2), \quad i = 1, ..., n; \qquad \epsilon_{z_j} \sim F(0, \sigma_{\epsilon_z}^2), \quad j = 1, ..., n_z.$$

Genetic correlation and heritability Given the above assumptions, we define the genetic correlation between $\boldsymbol{y}$ and $\boldsymbol{y}_z$ as

$$\varphi_{\beta\eta} = \frac{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\eta}}{\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}} \cdot \|\boldsymbol{\eta}\|_{\boldsymbol{\Sigma}}} \cdot \mathrm{I}(\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}} \cdot \|\boldsymbol{\eta}\|_{\boldsymbol{\Sigma}} > 0),$$

and we assume $\varphi_{\beta\eta} \in [-1, 1]$. Following Conditions 4.1 and 4.2, as $\min(n, n_z, m_{\beta\eta}) \to \infty$, the genetic correlation between $\boldsymbol{y}$ and $\boldsymbol{y}_z$ is asymptotically given by

$$\varphi_{\beta\eta} = \frac{m_{\beta\eta} \sigma_{\beta\eta} \mathrm{tr}(\boldsymbol{\Sigma})/p^2}{\{m_\beta \sigma_\beta^2 \mathrm{tr}(\boldsymbol{\Sigma})/p^2\}^{1/2} \{m_\eta \sigma_\eta^2 \mathrm{tr}(\boldsymbol{\Sigma})/p^2\}^{1/2}} = \kappa_{\beta\eta} \cdot \rho_{\beta\eta} + o_p(1).$$

Similarly, the heritability h_β^2 and h_η^2 defined in equation (4.2) can be asymptotically represented as

$$h_\beta^2 = \frac{\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\beta}\|_{\boldsymbol{\Sigma}} + \sigma_\epsilon^2} = \frac{m_\beta \sigma_\beta^2 \mathrm{tr}(\boldsymbol{\Sigma})/p^2}{m_\beta \sigma_\beta^2 \mathrm{tr}(\boldsymbol{\Sigma})/p^2 + \sigma_\epsilon^2} \quad \text{and} \quad h_\eta^2 = \frac{\|\boldsymbol{\eta}\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\eta}\|_{\boldsymbol{\Sigma}} + \sigma_{\epsilon_z}^2} = \frac{m_\eta \sigma_\eta^2 \mathrm{tr}(\boldsymbol{\Sigma})/p^2}{m_\eta \sigma_\eta^2 \mathrm{tr}(\boldsymbol{\Sigma})/p^2 + \sigma_{\epsilon_z}^2}.$$

With $\boldsymbol{\Sigma}_{ii} = 1$, $i = 1, \cdots, p$, we have $\mathrm{tr}(\boldsymbol{\Sigma})/p = 1$, and thus we have the same definitions of h_β^2 and h_η^2 as those in Jiang et al. (2016) and Guo et al. (2019) for the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

4.1.2 RMT lemmas

We introduce some known results from classic RMT (e.g., Tulino and Verdú (2004); Bai and Silverstein (2010); Paul and Aue (2014); Yao et al. (2015)) and some recent advances of trace functionals (e.g., Ledoit and Péché (2011); Dobriban and Wager (2018); Hastie et al. (2019); Wang et al. (2015)), which are foundations for our theoretical analysis of the large-scale GWAS data $(\boldsymbol{X}, \boldsymbol{Z})$. Below we mainly use the training data $\boldsymbol{X}$ as an example, but all the lemmas are applicable for the testing SNP data $\boldsymbol{Z}$ as well.

The ESD of $\widehat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ is given by $F_p^{\widehat{\boldsymbol{\Sigma}}}(x) = p^{-1}\sum_{i=1}^p \mathrm{I}\{\lambda_i(\widehat{\boldsymbol{\Sigma}}) \le x\}, x \in \mathbb{R}$. We are interested in the limit behavior of $F_p^{\widehat{\boldsymbol{\Sigma}}}(x)$, which has one-to-one correspondence with the limit behavior of its Stieltjes transform. For a general distribution $G(x)$ with support $I \subset \mathbb{R}$, the Stieltjes transform (e.g., page 514 of Bai and Silverstein (2010)) and its first order derivative (evaluated at z) are given by $s_G(z) = \int_{x\in I}(x-z)^{-1}dG(x)$ and $s_G'(z) = \int_{x\in I}(x-z)^{-2}dG(x)$, respectively, for $z \in \mathbb{C}\backslash I$. Therefore, let $I = [0, \infty)$, as $\min(n,p) \to \infty$, the Stieltjes transform of $F_p^{\widehat{\boldsymbol{\Sigma}}}(x)$ and its first order derivative are given by $s_{F_p}(z) = p^{-1}\mathrm{tr}\{(\widehat{\boldsymbol{\Sigma}} - z\boldsymbol{I}_p)^{-1}\}$ and $s_{F_p}'(z) = p^{-1}\mathrm{tr}\{(\widehat{\boldsymbol{\Sigma}} - z\boldsymbol{I}_p)^{-2}\}$, respectively, for $z \in \mathbb{C}\backslash I$. The asymptotic behavior of $F_p^{\widehat{\boldsymbol{\Sigma}}}(x)$ can be characterized in the following lemma (Marchenko and Pastur, 1967; Silverstein, 1995) by its Stieltjes transform. See, for example, Theorem 2.4 of Yao et al. (2015).

Lemma 4.1. *Under Condition 4.1, as $\min(n,p) \to \infty$, $F_p^{\widehat{\boldsymbol{\Sigma}}}(x)$ converges weakly to a limit probability distribution $M(x)$ with probability one, $x \in \mathbb{R}$. The Stieltjes transform of $M(x)$,*

donated as $s_M(z)$, is implicitly defined by the Marchenko-Pastur (M-P) equation

$$s_M(z) = \int \frac{1}{t\{1-\omega-\omega z s_M(z)\}-z} dH(t).$$

In general, $s_M(z)$ has no closed-form expression, but all information about the LSD $M(x)$ is contained in this equation. For the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have $s_M(z) = \{1-\omega-\omega z s_M(z)-z\}^{-1}$, it follows that (e.g., page 52 of Bai and Silverstein (2010))

$$s_M(z) = \frac{\sqrt{(1-\omega-z)^2-4\omega z}-(1-\omega-z)}{-2\omega z}, \quad \textit{and} \tag{4.3}$$

$$s_M^{'}(z) = \frac{(\omega-1)\sqrt{(-z-\omega+1)^2-4\omega z}-(\omega+1)z+(\omega-1)^2}{2\omega z^2\sqrt{(-z-\omega+1)^2-4\omega z}}, \tag{4.4}$$

respectively, for $z \in \mathbb{C}\backslash I$. Moreover, the LDS $M(x)$ in this special case is named the M-P law, and its probability density function is given by

$$p(x)_\omega = \frac{1}{2\pi\omega x} \cdot \sqrt{\big\{b_+(\omega)-x\big\}\big\{x-b_-(\omega)\big\}},$$

if $x \in [b_-(\omega), b_+(\omega)]$; and $p(x)_\omega = 0$ if $x \notin [b_-(\omega), b_+(\omega)]$; and $p(x)_\omega$ has a point mass $1-1/\omega$ at the origin if $\omega > 1$, where $b_\pm(\omega) = (1 \pm \sqrt{\omega})^2$.

Stieltjes transforms can also be used to study the limit of trace functionals of $F_p^{\widehat{\boldsymbol{\Sigma}}}(x)$. Below we donate $g(z) \equiv s_M(z)$ and $g^{'}(z) \equiv s_M^{'}(z)$, respectively. Sometimes, it is more convenient to use the notation defined on the $n \times n$ *companion* matrix $\widehat{\boldsymbol{\Phi}} = n^{-1}\boldsymbol{X}\boldsymbol{X}^T$ (Dobriban and Wager, 2018). Let $F_n^{\widehat{\boldsymbol{\Phi}}}(x) = n^{-1}\sum_{i=1}^n \mathrm{I}\big\{\lambda_i(\widehat{\boldsymbol{\Phi}}) \le x\big\}$ be the ESD of $\widehat{\boldsymbol{\Phi}}$, $x \in \mathbb{R}$, and let $P(x)$ be the limiting distribution of $F_n^{\widehat{\boldsymbol{\Phi}}}(x)$. Define the Stieltjes transform of $F_n^{\widehat{\boldsymbol{\Phi}}}(x)$ and its first order derivative as $s_{F_n}(z) = n^{-1}\mathrm{tr}\big\{\big(\widehat{\boldsymbol{\Phi}}-z\boldsymbol{I}_n\big)^{-1}\big\}$ and $s_{F_n}^{'}(z) = n^{-1}\mathrm{tr}\big\{\big(\widehat{\boldsymbol{\Phi}}-z\boldsymbol{I}_n\big)^{-2}\big\}$, respectively, and define the Stieltjes transform of $P(x)$ and its first order derivative as $v(z) \equiv s_P(z)$, and $v^{'}(z) \equiv s_P^{'}(z)$, respectively. We summarize the connections among $s_{F_p}(z)$, $s_{F_p}^{'}(z)$, $g(z)$, $g^{'}(z)$, $s_{F_n}(z)$, $s_{F_n}^{'}(z)$, $v(z)$ and $v^{'}(z)$ in the following lemma (Ledoit and Péché, 2011; Chen et al., 2015; Dobriban and Wager, 2018).

Lemma 4.2. *Under Condition 4.1, as min$(n,p)\to\infty$, for any $z\in\mathbb{C}\backslash I$, we have*

$$
\begin{aligned}
s_{F_p}(z) &= p^{-1}tr\big\{\big(\widehat{\boldsymbol{\Sigma}}-z\boldsymbol{I}_p\big)^{-1}\big\}\to_{a.s.} g(z),\\
\dot{s}_{F_p}(z) &= p^{-1}tr\big\{\big(\widehat{\boldsymbol{\Sigma}}-z\boldsymbol{I}_p\big)^{-2}\big\}\to_{a.s.} \dot{g}(z),\\
s_{F_n}(z) &= n^{-1}tr\big\{\big(\widehat{\boldsymbol{\Phi}}-z\boldsymbol{I}_n\big)^{-1}\big\}\to_{a.s.} v(z),\\
\dot{s}_{F_n}(z) &= n^{-1}tr\big\{\big(\widehat{\boldsymbol{\Phi}}-z\boldsymbol{I}_n\big)^{-2}\big\}\to_{a.s.} \dot{v}(z),\\
p^{-1}tr\big\{\boldsymbol{\Sigma}\big(\widehat{\boldsymbol{\Sigma}}-z\boldsymbol{I}_p\big)^{-1}\big\} &\to_{a.s.} \frac{1}{\omega}\Big\{\frac{1}{-zv(z)}-1\Big\},\\
p^{-1}tr\big\{\boldsymbol{\Sigma}\big(\widehat{\boldsymbol{\Sigma}}-z\boldsymbol{I}_p\big)^{-2}\big\} &\to_{a.s.} \frac{v(z)+z\dot{v}(z)}{\omega\big\{-zv(z)\big\}^2},\\
\omega\big\{g(z)+z^{-1}\big\} = v(z)+z^{-1},\quad &\textit{and}\quad \omega\big\{\dot{g}(z)-z^{-2}\big\} = \dot{v}(z)-z^{-2}.
\end{aligned}
$$

In general, little is known about the connection between population LSD $H(t)$ and empirical LDS $M(t)$. However, there is one-to-one correspondence between the moments of $H(t)$ and those of $M(t)$. For any positive integer k, define the kth moment of $H(t)$ as $b_k(\boldsymbol{\Sigma})=\int_{\mathbb{R}}t^k dH(t)=p^{-1}\mathrm{tr}(\boldsymbol{\Sigma}^k)$, and the kth moment of $M(t)$ as $b_k(\widehat{\boldsymbol{\Sigma}})=\int_{\mathbb{R}}t^k dM(t)=p^{-1}\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}^k)$. Then by Lemma 4.1, we have the following Lemma on the two sets of moments (Lemma 2.16 of Yao et al. (2015)).

Lemma 4.3. *Under Condition 4.1, as min$(n,p)\to\infty$, for any positive integer k, $b_k(\widehat{\boldsymbol{\Sigma}})$ is a function of $b_l(\boldsymbol{\Sigma})$, for $0<l\le k$, and ω. Specifically, the first three moments of $H(t)$ and the first three moments of $M(t)$ are linked as $b_1(\widehat{\boldsymbol{\Sigma}})=b_1(\boldsymbol{\Sigma})$, $b_2(\widehat{\boldsymbol{\Sigma}})=b_2(\boldsymbol{\Sigma})+\omega b_1(\boldsymbol{\Sigma})^2$, and $b_3(\widehat{\boldsymbol{\Sigma}})=b_3(\boldsymbol{\Sigma})+3\omega b_1(\boldsymbol{\Sigma})b_2(\boldsymbol{\Sigma})+\omega^2 b_1(\boldsymbol{\Sigma})^3$. Moreover, when $\boldsymbol{\Sigma}=\boldsymbol{I}_p$, we have $b_k(\boldsymbol{\Sigma})\equiv 1$ and $b_k(\widehat{\boldsymbol{\Sigma}})=\sum_{r=0}^{k-1}(r+1)^{-1}\binom{k}{r}\binom{k-1}{r}\omega^r$.*

For any positive integer k, since $\lambda_i(\boldsymbol{\Sigma})$ is uniformly bounded, $i=1,\cdots,p$, $b_k(\boldsymbol{\Sigma})$ and $b_k(\widehat{\boldsymbol{\Sigma}})$ are also bounded for any $\omega\in(0,\infty)$. Thus, we have the following lemma on the concentration of quadratic forms.

Lemma 4.4. *Under Condition 4.1, as min$(n,p)\to\infty$, for any positive integer k, we have $0<c\le b_k(\boldsymbol{\Sigma})\le b_k(\widehat{\boldsymbol{\Sigma}})\le C$. In addition, let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}=n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}=n_z^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$,*

$\boldsymbol{B}_{k_1,k_2} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{k_1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{k_2}$, *and define* $\boldsymbol{A}^0 = \boldsymbol{I}$ *for any matrix* $\boldsymbol{A}$. *Then for any non-negative integers* k_1, k_2, *we have*

$$\frac{tr\big(\boldsymbol{B}_{k_1,k_2}\boldsymbol{B}_{k_1,k_2}^T\big)}{\big\{tr(\boldsymbol{B}_{k_1,k_2})\big\}^2} = O(\frac{1}{p}) = o(1).$$

Moreover, let $\boldsymbol{\alpha}$ *be a p-dimensional random vector of i.i.d. elements with mean zero, variance* σ_α^2, *and finite fourth order moment, we have*

$$\boldsymbol{\alpha}^T\boldsymbol{B}_{k_1,k_2}\boldsymbol{\alpha} = \sigma_\alpha^2 \cdot tr(\boldsymbol{B}_{k_1,k_2}) \cdot \{1 + o_p(1)\}.$$

The proof of Lemma 4.4 can be found in Appendix B, which is based on the Lemma B.26 of Bai and Silverstein (2010) and the Markov's inequality. Lemma 4.4 shows that the quadratic forms of $\boldsymbol{B}_{k_1,k_2}$ concentrate around their means. We note that $\omega \in (0, \infty)$ is a key condition. When $\omega = \infty$, the concentration still holds when either k_1 or k_2 is zero, but may not hold when both k_1 and k_2 are nonzero.

4.2 Marginal estimator

Let $\widehat{\boldsymbol{\beta}}$ be a generic $p \times 1$ estimator of $\boldsymbol{\beta}$, the out-of-sample predictor and in-sample estimation are given by $\widehat{\boldsymbol{S}}_{\boldsymbol{Z}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{S}}_{\boldsymbol{X}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, respectively. The out-of-sample and in-sample R^2 are, respectively, defined as A^2 and E^2, where

$$A = \frac{\boldsymbol{y}_z^T\widehat{\boldsymbol{S}}_{\boldsymbol{Z}}}{\|\boldsymbol{y}_z\| \cdot \|\widehat{\boldsymbol{S}}_{\boldsymbol{Z}}\|} \quad \text{and} \quad E = \frac{\boldsymbol{y}^T\widehat{\boldsymbol{S}}_{\boldsymbol{X}}}{\|\boldsymbol{y}\| \cdot \|\widehat{\boldsymbol{S}}_{\boldsymbol{X}}\|}. \tag{4.5}$$

In this section, we present the results of A^2 and E^2 for marginal estimator $\widehat{\boldsymbol{\beta}}_S$, donated as A_S^2 and E_S^2, respectively.

4.2.1 Asymptotic limits

The asymptotic limits of A_S^2 and E_S^2 are given in the following theorem.

Theorem 4.1. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*(n, n_z, $m_{\beta\eta}$,

$p) \to \infty$, *for any* $\omega \in (0, \infty)$, $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, *and* $\boldsymbol{\Sigma}$, *we have*

$$\begin{aligned} A_S^2 &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{n\{tr(\widehat{\boldsymbol{\Sigma}}_X \widehat{\boldsymbol{\Sigma}}_Z)\}^2 \cdot h_\beta^2}{n tr(\widehat{\boldsymbol{\Sigma}}_Z) tr(\widehat{\boldsymbol{\Sigma}}_X \widehat{\boldsymbol{\Sigma}}_Z \widehat{\boldsymbol{\Sigma}}_X) \cdot h_\beta^2 + tr(\widehat{\boldsymbol{\Sigma}}_Z) tr(\widehat{\boldsymbol{\Sigma}}_X) tr(\widehat{\boldsymbol{\Sigma}}_Z \widehat{\boldsymbol{\Sigma}}_X) \cdot (1 - h_\beta^2)} + o_p(1) \\ &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \left\{ \frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \frac{\omega}{b_2(\boldsymbol{\Sigma})} \cdot \frac{1}{h_\beta^2} \right\}^{-1} + o_p(1), \end{aligned}$$

and

$$\begin{aligned} E_S^2 &= \frac{\{n tr(\widehat{\boldsymbol{\Sigma}}_X^2) \cdot h_\beta^2 + tr(\widehat{\boldsymbol{\Sigma}}_X)^2 \cdot (1 - h_\beta^2)\}^2}{n^2 tr(\widehat{\boldsymbol{\Sigma}}_X) tr(\widehat{\boldsymbol{\Sigma}}_X^3) \cdot h_\beta^2 + n tr(\widehat{\boldsymbol{\Sigma}}_X)^2 tr(\widehat{\boldsymbol{\Sigma}}_X^2) \cdot (1 - h_\beta^2)} + o_p(1) \\ &= \frac{\{b_2(\boldsymbol{\Sigma}) \cdot h_\beta^2 + \omega\}^2}{\{b_2(\boldsymbol{\Sigma}) \cdot h_\beta^2 + \omega\}^2 + b_2(\boldsymbol{\Sigma})\omega + \{b_3(\boldsymbol{\Sigma}) - b_2(\boldsymbol{\Sigma})^2 \cdot h_\beta^2\} \cdot h_\beta^2} + o_p(1), \end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_X = n^{-1} \boldsymbol{X}^T \boldsymbol{X}$, $\widehat{\boldsymbol{\Sigma}}_Z = n_z^{-1} \boldsymbol{Z}^T \boldsymbol{Z}$, *and* $b_k(\boldsymbol{\Sigma}) = \int_{\mathbb{R}} t^k dH(t) = p^{-1} tr(\boldsymbol{\Sigma}^k)$ *for* $k = 1, 2,$ *and* 3. *For* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, *we have* $b_k(\boldsymbol{\Sigma}) = 1$ *for any positive integral* k,

$$A_S^2 = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{h_\beta^2}{h_\beta^2 + \omega} + o_p(1), \quad \textit{and} \quad E_S^2 = \frac{(h_\beta^2 + \omega)^2}{(h_\beta^2 + \omega)^2 + \omega + h_\beta^2(1 - h_\beta^2)} + o_p(1). \tag{4.6}$$

Theorem 4.1 has several important implications. First, h_η^2 is the overall genetic effects in $\boldsymbol{y}_z$ and $\varphi_{\beta\eta}^2$ represents the genetic similarity between training and testing data. Thus, $h_\eta^2 \varphi_{\beta\eta}^2$ can be viewed as the signal strength of out-of-sample cross-trait prediction and also the upper bound of prediction performance. We find that the gap between prediction accuracy A_S^2 and the upper bound $h_\eta^2 \varphi_{\beta\eta}^2$ is determined by a function of $b_3(\boldsymbol{\Sigma})$, $b_2(\boldsymbol{\Sigma})$, ω, and h_β^2. When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, equation (4.6) indicates that A_S^2 is linearly decayed away from $h_\eta^2 \varphi_{\beta\eta}^2$ by the nonzero ω. It is also easy to see that the gap disappears when $\omega = 0$. For in-sample R^2, we have $E_S^2 = h_\beta^2$ if $\omega = 0$, but in general it does not hold for nonzero ω. These results clearly illustrate the difference between low- and high-dimensions, and quantify the gap between true signal strength and out-of-sample/in-sample performance of high-dimensional marginal estimator.

Second, there are bidirectional influences of feature-wise correlation for general correlation

$\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$. Specifically, A_S^2 depends on the first three moments of the LSD $H(t)$ of $\boldsymbol{\Sigma}$ through the two terms $b_2(\boldsymbol{\Sigma})^2/b_3(\boldsymbol{\Sigma})$ and $\omega/b_2(\boldsymbol{\Sigma})$. It follows from by Cauchy–Schwarz inequality that $b_3(\boldsymbol{\Sigma}) > b_2(\boldsymbol{\Sigma})^2$ and $b_2(\boldsymbol{\Sigma}) > 1$ hold. In a classic linear model setting where p is fixed (or $\omega = 0$), we have $A_S^2 = h_\eta^2 \varphi_{\beta\eta}^2 \cdot b_2(\boldsymbol{\Sigma})^2/b_3(\boldsymbol{\Sigma})$. Thus, A_S^2 is reduced by a factor $b_2(\boldsymbol{\Sigma})^2/b_3(\boldsymbol{\Sigma})$ due to the unadjusted feature-wise correlation. On the other hand, when $\omega > 0$, further decay of A_S^2 is introduced by the nonzero term $\omega/b_2(\boldsymbol{\Sigma})$. Since $b_2(\boldsymbol{\Sigma}) > 1$ holds, correlation among features can delay this type of decay. This makes sense, because correlation among p features can be regarded as a reduction of signal dispersion in high-dimensions. Together, there is a transition point for whether or not feature-wise correlation can help achieve higher prediction accuracy in high-dimensions. Formally, we can define the *prediction relative efficiency* (PRE) for $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$ to quantify the bidirectional effects of $\boldsymbol{\Sigma}$ on A_S^2 and identify the transition point. Let $\delta_S(\boldsymbol{\Sigma}) = A_S^2(\boldsymbol{\Sigma})/A_S^2(\boldsymbol{I}_p)$, we have

$$\delta_S(\boldsymbol{\Sigma}) = \frac{h_\beta^2 + \omega}{h_\beta^2 \cdot \frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \omega \cdot \frac{1}{b_2(\boldsymbol{\Sigma})}} + o_p(1),$$

and it follows that

$$\begin{array}{ccc} > & & > \\ \delta_S(\boldsymbol{\Sigma}) = 1 + o_p(1) & \text{if} & \omega = h_\beta^2 \cdot \dfrac{b_3(\boldsymbol{\Sigma}) - b_2(\boldsymbol{\Sigma})^2}{b_2(\boldsymbol{\Sigma})^2 - b_2(\boldsymbol{\Sigma})}. \\ < & & < \end{array}$$

Third, we consider the difference between in-sample and out-of-sample R^2. For the optimal case with $h_\beta^2 = h_\eta^2 = \varphi_{\beta\eta}^2 = 1$, we have

$$A_S^2 = \frac{\left\{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_X \widehat{\boldsymbol{\Sigma}}_Z)\right\}^2}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_Z)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_Z \widehat{\boldsymbol{\Sigma}}_X^2)} + o_p(1) \quad \text{and} \quad E_S^2 = \frac{\left\{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_X^2)\right\}^2}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_X)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_X^3)} + o_p(1).$$

This optimal case reveals more insights into the difference between in-sample and out-of-sample R^2, and the difference between low- and high-dimensions. In low-dimension with

$\omega = 0$, in-sample and out-of-sample R^2 have the same limit

$$A_S^2 = E_S^2 = \frac{\left\{\mathrm{tr}(\mathbf{\Sigma}^2)\right\}^2}{\mathrm{tr}(\mathbf{\Sigma}) \cdot \mathrm{tr}(\mathbf{\Sigma}^3)} + o_p(1).$$

For nonzero ω, we note that $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X) = \mathrm{tr}(\widehat{\mathbf{\Sigma}}_Z) = \mathrm{tr}(\mathbf{\Sigma})$, and $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X\widehat{\mathbf{\Sigma}}_Z) = \mathrm{tr}(\mathbf{\Sigma}^2)$ for any ω. That is, trace of sample covariance is (asymptotically) the same as the trace of population covariance, and similar result holds for the product of two independent sample covariances. However, by Lemma 4.3, this kind of concordance no longer holds for the trace of higher order products in high-dimensions with nonzero ω. Specifically, we have $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X^2) = \mathrm{tr}(\mathbf{\Sigma}^2) + n^{-1}\mathrm{tr}(\mathbf{\Sigma})^2$, $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X^3) = \mathrm{tr}(\mathbf{\Sigma}^3) + 3n^{-1}\mathrm{tr}(\mathbf{\Sigma})\mathrm{tr}(\mathbf{\Sigma}^2) + n^{-2}\mathrm{tr}(\mathbf{\Sigma})^3$, and $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_Z\widehat{\mathbf{\Sigma}}_X^2) = \mathrm{tr}(\mathbf{\Sigma}\widehat{\mathbf{\Sigma}}_X^2)$ $= \mathrm{tr}(\mathbf{\Sigma}^3) + n^{-1}\mathrm{tr}(\mathbf{\Sigma})\mathrm{tr}(\mathbf{\Sigma}^2)$. It follows that

$$A_S^2 = \frac{\left\{\mathrm{tr}(\mathbf{\Sigma}^2)\right\}^2}{\mathrm{tr}(\mathbf{\Sigma}) \cdot \left\{\mathrm{tr}(\mathbf{\Sigma}^3) + n^{-1}\mathrm{tr}(\mathbf{\Sigma})\mathrm{tr}(\mathbf{\Sigma}^2)\right\}} + o_p(1)$$

and

$$E_S^2 = \frac{\left\{\mathrm{tr}(\mathbf{\Sigma}^2) + n^{-1}\mathrm{tr}(\mathbf{\Sigma})^2\right\}^2}{\mathrm{tr}(\mathbf{\Sigma}) \cdot \left\{\mathrm{tr}(\mathbf{\Sigma}^3) + 3n^{-1}\mathrm{tr}(\mathbf{\Sigma})\mathrm{tr}(\mathbf{\Sigma}^2) + n^{-2}\mathrm{tr}(\mathbf{\Sigma})^3\right\}} + o_p(1).$$

Therefore, due to the different trace limits, A_S^2 and E_S^2 can become completely different as ω increases.

In summary, the asymptotic performance of high-dimensional marginal estimator is solely determined by heritability, genetic correlation, ω, and the first three moments of $H(t)$. These parameters are independent from the unknown numbers m_β, m_η, and $m_{\beta\eta}$. Such properties enable us to easily evaluate the prediction accuracy of a given GWAS dataset regardless of the underlying number of true signals. In addition, PRE measures the influence of $\mathbf{\Sigma}$ on A_S^2, which can also be used to compare the prediction accuracy among different structures of $\mathbf{\Sigma}$. In next section, we illustrate how to apply the Theorem 4.1 to estimate A_S^2 in GWAS applications.

4.2.2 Prediction accuracy estimation and comparison

In GWAS, different global populations (e.g., African, Latino, East Asian) have different SNP correlation structure $\mathbf{\Sigma}$, and $\mathbf{\Sigma}$ is known to be largely consistent within each population (Gurdasani et al., 2019). Thus, given the same ω and $h_\eta^2, h_\beta^2, \varphi_{\beta\eta}^2$, the prediction accuracy of GWAS data varies across different populations. To evaluate and compare the prediction accuracy of GWAS data in diverse populations, we need to study the LSD $H(x)$ of $\mathbf{\Sigma}$. Here we discuss two approaches to evaluate the prediction accuracy A_S^2 for each global population.

Asymptotic estimator (External reference panel) The asymptotic estimator is based on the asymptotic limits. It is clear that we only need to estimate the first three moments $b_1(\mathbf{\Sigma})$, $b_2(\mathbf{\Sigma})$, $b_3(\mathbf{\Sigma})$ of $H(t)$, which have known relationships with $b_1(\widehat{\mathbf{\Sigma}})$, $b_2(\widehat{\mathbf{\Sigma}})$, $b_3(\widehat{\mathbf{\Sigma}})$ according to Lemma 4.3. Therefore, we can estimate $b_k(\widehat{\mathbf{\Sigma}})$ from SNP data then obatin $b_k(\mathbf{\Sigma})$, for $k = 1, 2, 3$. In practice, this can be done using external data in publicly available LD reference panels (Tam et al., 2019), such as the 1000 Genomes Project (1000-Genomes-Project-Consortium., 2015). Let the reference data be $\boldsymbol{W} \in R^{n_w \times p}$, and let $\widehat{\mathbf{\Sigma}}_W = n_w^{-1}\boldsymbol{W}^T\boldsymbol{W}$, then $b_k(\widehat{\mathbf{\Sigma}}_W) = p^{-1}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_W^k) = p^{-1}\sum_{i=1}^p \lambda_i(\widehat{\mathbf{\Sigma}}_W)^k$, $k = 1, 2, 3$. Thus, all we need are the eigenvalues of $\widehat{\mathbf{\Sigma}}_W$, $\lambda_i(\widehat{\mathbf{\Sigma}}_W)$, $i = 1, \cdots, p$. When $n_w < p$, we may instead focus on the $n \times n$ companion matrix $\widehat{\mathbf{\Phi}}_W = n_w^{-1}\boldsymbol{W}\boldsymbol{W}^T$ to obtain these moments, since the nonzero eigenvalues of $\widehat{\mathbf{\Phi}}_W$ and $\widehat{\mathbf{\Sigma}}_W$ are the same.

Empirical estimator (Individual-level data) When SNP data $\boldsymbol{X}$ and $\boldsymbol{Z}$ are available, one can also directly estimate the prediction accuracy by evaluating the four traces $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X)$, $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_Z)$, $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X\widehat{\mathbf{\Sigma}}_Z)$, and $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X^2\widehat{\mathbf{\Sigma}}_Z)$. Since $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X) = \mathrm{tr}(\widehat{\mathbf{\Sigma}}_Z) = p$, we only need to estimate $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X\widehat{\mathbf{\Sigma}}_Z)$ and $\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X^2\widehat{\mathbf{\Sigma}}_Z)$. Estimating $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_Z$ can be computationally expensive when both n and p are large. However, some tools have been developed to tackle this challenge (Quick et al., 2018; Das et al., 2016). Moreover, we may need to additionally account for the population stratification when population substructures exist (Sun and Lin, 2017). One common solution is to remove the top few "outlier" eigenvalues, which often represent population substructures

if any, since the population substructures are usually much stronger than the local SNP correlations. Next, we discuss two more potential usages of the prediction accuracy results.

Diverse populations Based on these estimators, one can compare the prediction accuracy among diverse populations using their PREs. For example, suppose population 1 has $\boldsymbol{\Sigma}_1$ and population 2 has $\boldsymbol{\Sigma}_2$, then their relative prediction accuracy can be written by the ratio of their PREs

$$\frac{\delta_S(\boldsymbol{\Sigma}_1)}{\delta_S(\boldsymbol{\Sigma}_2)} = \frac{h_\beta^2 \cdot \frac{b_3(\boldsymbol{\Sigma}_2)}{b_2(\boldsymbol{\Sigma}_2)^2} + \omega \cdot \frac{1}{b_2(\boldsymbol{\Sigma}_2)}}{h_\beta^2 \cdot \frac{b_3(\boldsymbol{\Sigma}_1)}{b_2(\boldsymbol{\Sigma}_1)^2} + \omega \cdot \frac{1}{b_2(\boldsymbol{\Sigma}_1)}} + o_p(1).$$

It is clear that when ω is much larger than h_β^2 and $b_3(\boldsymbol{\Sigma})$ is comparable to $b_2(\boldsymbol{\Sigma})^2$, $b_2(\boldsymbol{\Sigma})$ plays an important role in the relative prediction accuracy.

LD-based pruning In practice, it is quite common to first perform LD-based pruning with predefined threshold to remove highly related SNPs (e.g., remove one of a pair of SNPs that have correlation larger than the threshold) before out-of-sample prediction. The choice of the predefined threshold is often arbitrary. Using Theorem 4.1, it is possible to input a series of thresholds, estimate the corresponding prediction accuracy, and then make a decision about the "optimal" threshold for SNP pruning.

4.2.3 Meta-analysis of marginal estimator

Motivated by GWAS applications, we extend our analysis to consider meta-analysis of multiple marginal estimators from $k \in (0, \infty)$ independent GWAS $\{(\boldsymbol{X}_i, \boldsymbol{y}_i) : i = 1, \ldots, k\}$ on the same trait with genetic effects $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1^T, \ldots, \widehat{\boldsymbol{\beta}}_k^T)$ be a $p \times k$ matrix of marginal estimators from all k GWAS. Let $\boldsymbol{d} = (d_i, \ldots, d_k)^T$ be a $k \times 1$ vector of weights, and let $\widehat{\boldsymbol{B}}(\boldsymbol{d}) = \widehat{\boldsymbol{B}}\boldsymbol{d}$ be the aggregated summary statistics. We also denote $A_S^2(\boldsymbol{d})$ and $E_S^2(\boldsymbol{d})$ as the out-of-sample and in-sample R^2 for marginal estimator $\widehat{\boldsymbol{B}}(\boldsymbol{d})$, respectively. We have the following results from the k studies.

Corollary 4.1. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, suppose we have independent GWAS* $(\boldsymbol{X}_i, \boldsymbol{y}_i)$ *with sample sizes* n_i *and* p *SNPs for* $i = 1, \ldots, k$. *As min*$(n_1, \ldots, n_k, n_z, m_{\beta\eta}, p) \to \infty$, *for any* $\omega \in (0, \infty)$, $h^2_\beta, h^2_\eta \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, $\boldsymbol{d}$, *and* $\boldsymbol{\Sigma}$, *we have*

$$A^2_S(\boldsymbol{d}) = h^2_\eta \varphi^2_{\beta\eta} \cdot \left\{ \frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \frac{\sum_{i=1}^k d_i^2/n_i}{(\sum_{i=1}^k d_i)^2} \cdot \frac{p}{b_2(\boldsymbol{\Sigma}) h^2_\beta} \right\}^{-1} + o_p(1).$$

For $\boldsymbol{d} \equiv \boldsymbol{d}^* = (n_1, \ldots, n_k)^T$, *we have*

$$A^2_S(\boldsymbol{d}^*) = h^2_\eta \varphi^2_{\beta\eta} \cdot \left\{ \frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \frac{\widetilde{\omega}_k}{h^2_\beta} \cdot \frac{1}{b_2(\boldsymbol{\Sigma})} \right\}^{-1} + o_p(1),$$

where $\widetilde{\omega}_k = p/\sum_{i=1}^k n_i$. *The* $A^2_S(\boldsymbol{d}^*)$ *is the same as the out-of-sample* R^2 *for one single GWAS with sample size* $\sum_{i=1}^k n_i$. *Particularly, when* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, *we have*

$$A^2_S(\boldsymbol{d}^*) = h^2_\eta \varphi^2_{\beta\eta} \cdot \frac{h^2_\beta}{h^2_\beta + \widetilde{\omega}_k} + o_p(1).$$

Corollary 4.1 shows that marginal screening has no prediction accuracy loss in distributed computing followed by meta-analysis with weights in $\boldsymbol{d}^*$. Thus, aggregating summary statistics from independent training GWAS has the same asymptotic prediction accuracy as one big GWAS that trains all the individual-level data together. This is a favorable property of high-dimensional marginal estimator. It is known that both OLS (Dobriban and Sheng, 2018) and ridge (Dobriban and Sheng, 2019) estimators may have prediction accuracy loss in high-dimensional distributed computation. Similar results also hold for in-sample R^2. For example, when $\boldsymbol{d} = \boldsymbol{d}^*$, we have

$$E^2_S(\boldsymbol{d}^*) = \frac{\{b_2(\boldsymbol{\Sigma}) \cdot h^2_\beta + \widetilde{\omega}_k\}^2}{\{b_2(\boldsymbol{\Sigma}) \cdot h^2_\beta + \widetilde{\omega}_k\}^2 + b_2(\boldsymbol{\Sigma})\widetilde{\omega}_k + \{b_3(\boldsymbol{\Sigma}) - b_2(\boldsymbol{\Sigma})^2 \cdot h^2_\beta\} \cdot h^2_\beta} + o_p(1).$$

4.3 The class of ridge-type estimators

In this section, we present the results for the following ridge-type estimators: $\{\widehat{\boldsymbol{\beta}}_R(\lambda)$, $\widehat{\boldsymbol{\beta}}_B(\tau)$, $\widehat{\boldsymbol{\beta}}_R(0^+)$, $\widehat{\boldsymbol{\beta}}_O\}$. We define their out-of-sample R^2 as $\{A_R^2(\lambda)$, $A_B^2(\tau)$, $A_R^2(0^+)$, $A_O^2\}$ and in-sample R^2 as $\{E_R^2(\lambda), E_B^2(\tau), E_R^2(0^+), E_O^2\}$, respectively.

4.3.1 Out-of-sample *R*-squared

We have the following results on $\{A_R^2(\lambda)$, $A_B^2(\tau)$, $A_R^2(0^+)$, $A_O^2\}$.

Theorem 4.2. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n, n_z, m_{\beta\eta}, p) \to \infty$*, for any* $\omega \in (0, \infty)$, $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, *and* $\boldsymbol{\Sigma}$, *we have*

$$\begin{aligned} A_R^2(\lambda) &= A_B^2(\lambda/\omega) \\ &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\left[1 + \frac{\lambda}{\omega}\{1 - \frac{1}{\lambda v(-\lambda)}\}\right]^2 \cdot h_\beta^2}{\left[1 + \frac{\lambda}{\omega}\{2 - \frac{1}{\lambda v(-\lambda)} - \frac{\dot{v}(-\lambda)}{v(-\lambda)^2}\}\right] \cdot h_\beta^2 + \{\frac{\dot{v}(-\lambda)}{v(-\lambda)^2} - 1\} \cdot (1 - h_\beta^2)} + o_p(1), \end{aligned}$$

and

$$A_R^2(0^+) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\{1 - \frac{1}{v(0^+)\omega}\}^2 \cdot h_\beta^2}{\{1 - \frac{1}{v(0^+)\omega}\} \cdot h_\beta^2 + \{\frac{\dot{v}(0^+)}{v(0^+)^2} - 1\} \cdot (1 - h_\beta^2)} + o_p(1),$$

where $v(0^+) = \lim_{\lambda \to 0^+} v(-\lambda)$ *and* $\dot{v}(0^+) = \lim_{\lambda \to 0^+} \dot{v}(-\lambda)$. *Here* $v(-\lambda)$ *is the Stieltjes transform related to* $\boldsymbol{\Sigma}$ *and* $\dot{v}(-\lambda)$ *is its first order derivative. The* $A_R^2(0^+)$ *reduces to* A_O^2 *if* $\omega < 1$, *which is given by*

$$A_O^2 = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \left\{1 + \frac{1 - h_\beta^2}{h_\beta^2} \cdot \frac{\omega}{1 - \omega}\right\}^{-1} + o_p(1).$$

If $h_\beta^2 \in (0, 1)$, *then* $A_R^2(\lambda)$ *is maximized at* $\lambda = \lambda^* \equiv \omega \cdot (1 - h_\beta^2)/h_\beta^2$, *and the optimal out-of-sample* R^2 *is given by*

$$A_R^2(\lambda^*) = A_B^2(\lambda^*/\omega) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \left\{\frac{1}{h_\beta^2} - \frac{1}{v(-\lambda^*)\omega}\right\} + o_p(1).$$

If $h_\beta^2 = 1$, i.e., $\boldsymbol{y}$ is a fully heritable trait, then the optimal out-of-sample R^2 is obtained as $\lambda \to 0^+$, and we have

$$A_R^2(0^+) = \begin{cases} h_\eta^2 \varphi_{\beta\eta}^2 + o_p(1), & \text{if} \quad \omega < 1; \\ h_\eta^2 \varphi_{\beta\eta}^2 \cdot \left\{1 - \frac{1}{v(0^+)\omega}\right\} + o_p(1), & \text{if} \quad \omega > 1. \end{cases}$$

Theorem 4.2 shows how the out-of-sample prediction accuracy decays away from the true signal strength $h_\eta^2 \varphi_{\beta\eta}^2$. The A_O^2 is invariant to $\boldsymbol{\Sigma}$ and always has a closed-form expression, and thus the gap is only determined by h_β^2 and ω. For all other estimators, due to the linear shrinkage induced by nonzero λ, $\boldsymbol{\Sigma}$ still has influence on the gap through the limits of Stieltjes transform $v(-\lambda)$ and its first order derivative $\dot{v}(-\lambda)$. When $\lambda = \lambda^*$, $\dot{v}(-\lambda)$ cancels out and thus the optimal out-of-sample R^2 depends on $\boldsymbol{\Sigma}$ only through $v(-\lambda)$. Let $\mathrm{STN}(h_\beta^2) = h_\beta^2/(1 - h_\beta^2)$ be the signal to noise ratio, λ^* can be rewritten as $\lambda^* = \omega/\mathrm{STN}(h_\beta^2)$. In other words, for the linear shrinkage estimator $\widehat{\boldsymbol{\Sigma}}_X + \lambda \boldsymbol{I}_p$, the optimal weight for $\boldsymbol{I}_p$ is proportional to ω and inversely proportional to the signal to noise ratio, matching similar results on mean squared prediction error (Dobriban and Wager, 2018).

When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, the closed-form expressions for $v(-\lambda)$ and $\dot{v}(-\lambda)$ are available, and thus we have closed-form expressions for $\left\{A_R^2(\lambda),\, A_B^2(\tau),\, A_R^2(0^+),\, A_O^2\right\}$. In Appendix B, we further quantify the relative prediction accuracy between marginal estimator and the optimal ridge estimator in closed-forms. We provide more insights into high-dimensional dense signal prediction as follows.

The first one is on optimal regularizer and prediction accuracy. The optimal values λ^* and τ^* for out-of-sample R^2 (when $h_\beta^2 \in (0, 1)$) are functions of h_β^2 and ω, and are independnet of all other parameters including $\boldsymbol{\Sigma}$. Since ω is known and consistent estimator of h_β^2 is available in GWAS context (Yang et al., 2010; Jiang et al., 2016; Ma and Dicker, 2019), cross-validation techniques are not required to obtain optimal regularizers for ridge or BLUP estimators. We provide a brief introduction and discussion of common GWAS estimators of h_β^2 in Appendix B. Moreover, the asymptotic prediction accuracy $A_R^2(\lambda^*)$ can

be estimated by additionally calculating $s_{F_n}(\lambda^*)$ (see Appendix B), which is a consistent estimator of $v(-\lambda^*, \boldsymbol{\Sigma})$. To quantify the influence of $\boldsymbol{\Sigma}$ on $A_R^2(\lambda^*)$, we define PRE for $\boldsymbol{\Sigma} \neq \boldsymbol{I}_p$ as $\delta_R(\lambda^*, \boldsymbol{\Sigma}) = A_R^2(\lambda^*, \boldsymbol{\Sigma})/A_R^2(\lambda^*, \boldsymbol{I}_p)$, and then we have

$$\delta_R(\lambda^*, \boldsymbol{\Sigma}) = \frac{\omega + h_\beta^2 - \sqrt{(\omega - h_\beta^2)^2 + 4\omega h_\beta^2(1 - h_\beta^2)}}{2\omega - 2h_\beta^2 / v(-\lambda^*, \boldsymbol{\Sigma})} + o_p(1).$$

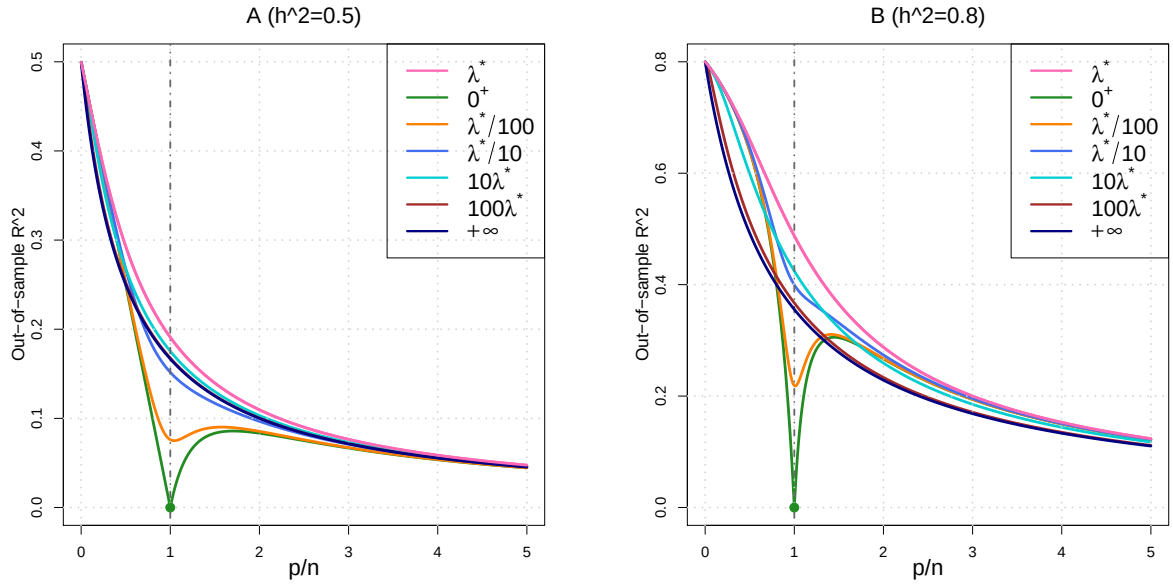

Figure 4.1: Out-of-sample R-squared $A_R^2(\lambda)$ of ridge-type estimators given different λ and heritability when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. λ^* is the optimal λ value, 0^+ corresponds to the case $\lambda \to 0^+$, and $+\infty$ represents $\lambda \to +\infty$. We set $\varphi_{\beta\eta} = 1$, $h_\beta^2 = h_\eta^2 = (0.5, 0.8)$ in A and B, respectively.

The second one is on over- and under-fitting issues. Figure 4.1 displays $A_R^2(\lambda)$ across different λ, ω, and h^2. It is clear that $A_R^2(\lambda)$ is near-optimal for any λ when ω is big (e.g., $\omega > 5$), especially when h^2 is not high. In contrast, when $\omega \approx 1$, model over-fitting with small λ should be avoided and model under-fitting (over-regularization) with large λ can be substantially sub-optimal. Notably, $A_R^2(0^+)$ can become surprisingly small. For $\omega > 1$, $A_R^2(0^+)$ is not a monotone function of ω, and the optimal value is achieved at $\omega = 1 + \sqrt{1 - h_\beta^2}$. When ω decreases from $1 + \sqrt{1 - h_\beta^2}$ towards 1, $A_R^2(0^+)$ reduces dramatically.

The third one is on the blessing of dimensionality. When ω is large, $A_R^2(\lambda)$ has almost

identical performance for all λ. Particularly, the out-of-sample R^2 of $\widehat{\boldsymbol{\beta}}_S$ is similar to that of $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$, then $\widehat{\boldsymbol{\beta}}_S$ can be a good choice for out-of-sample applications because it is much more computationally efficient than $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$. High dimensionality indeed reduces the required computational burden because marginal and conditional estimators yield similar out-of-sample performance, which is quite counterintuitive.

The fourth one is on the curse of dimensionality. The upper limit of the prediction accuracy of all ridge-type estimators might be not satisfactory when ω is large. For example, when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, consider the optimal case where $h_\eta^2 = h_\beta^2 = \varphi_{\beta\eta}^2 = 1$, the asymptotic optimal out-of-sample R^2 of ridge-type estimators is one for $n > p$. However, when $n < p$, the out-of-sample R^2 has a upper bound of $n/p = 1/\omega$, which can be viewed as the ratio of sample size and model complexity. In addition, $1/\omega$ can hardly be achieved in practical situations if any. To see this, consider $h_\eta^2 = h_\beta^2 \in (0, 1)$ and $\varphi_{\beta\eta}^2 = 1$, we can rewrite the optimal out-of-sample R^2 as

$$A_R^2(\lambda^*) = A_B^2(\lambda^*/\omega) = \frac{p + nh_\beta^2 - |p - nh_\beta^2| \cdot \sqrt{1 + \Delta}}{2p} + o_p(1),$$

where $\Delta = 4h_\beta^2(1 - h_\beta^2)/(\omega^{1/2} - \omega^{-1/2}h_\beta^2)^2 > 0$. Since $\Delta \approx 0$ only holds for large ω and $h_\beta^2 \approx 1$, $A_R^2(\lambda^*)$ is close to the upper bound $n/p \cdot h_\beta^2$ only when ω is large for highly heritable traits prediction (Figure 4.2). For general $\boldsymbol{\Sigma}$, similar to the discussions on marginal estimator, feature-wise correlation can delay the negative influences of growing dimension, but the general pattern remains the same. This finding reveals a fundamental challenge in high-dimensional dense signal prediction.

The fifth one is on the unboundedness of $\mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}$. The $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$ has better out-of-sample R^2 than $\widehat{\boldsymbol{\beta}}_O$ for $\omega \in (0, 1)$. As shown in Figure 4.2, when ω is close to zero, $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$ and $\widehat{\boldsymbol{\beta}}_O$ are close to each other, which matches the classic results in linear models. However, as ω moves towards one, the performance of $\widehat{\boldsymbol{\beta}}_O$ is much worse than $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$. One way to explain this surprising behavior of $\widehat{\boldsymbol{\beta}}_O$ is that $\mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}$ can become very large when

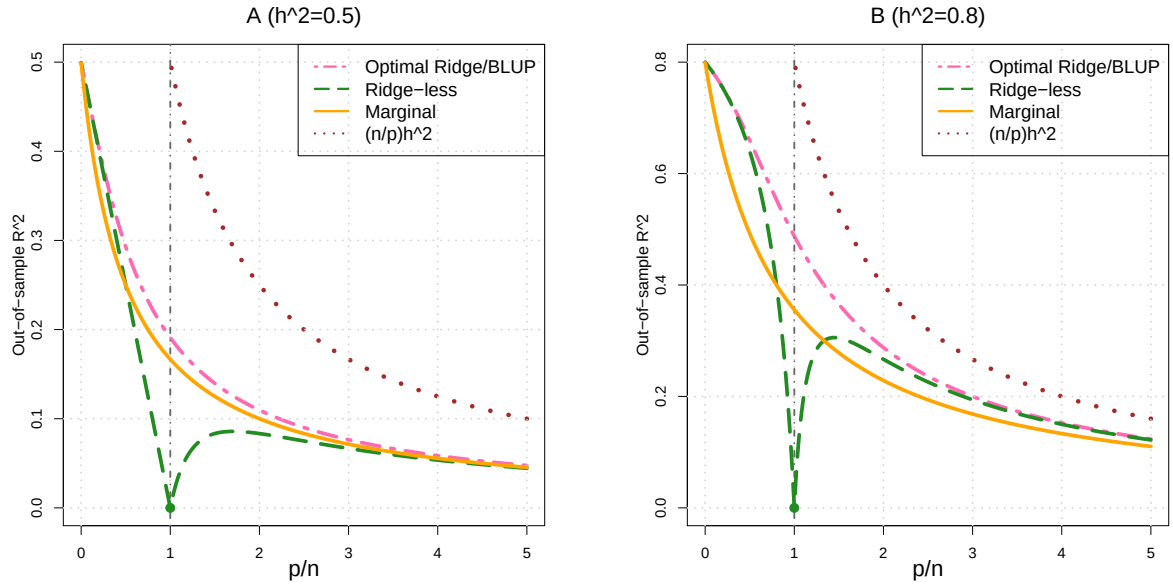

Figure 4.2: Out-of-sample R-squared of optimal ridge/BLUP estimators ($A_R^2(\lambda^*) = A_B^2(\lambda^*/\omega)$), ridge/BLUP-less estimators ($A_R^2(0^+) = A_B^2(0^+)$), and marginal estimator (A_S^2) when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. We set $\varphi_{\beta\eta} = 1$, $h_\beta^2 = h_\eta^2 = (0.5, 0.8)$ in A and B, respectively. The dash line represents the upper limit $(n/p) \cdot h_\beta^2$.

$\omega \to 1^-$. To see this, consider $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, when $\omega < 1$, we have

$$A_O^2 = A_R^2(0) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{h_\beta^2}{h_\beta^2 + \mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\} \cdot (1 - h_\beta^2)} + o_p(1).$$

In Gaussian case, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ follows the inverse Wishart distribution and the mean of $\mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}$ is $\omega/(1-\omega-1/n)$, which can be large as $\omega \to 1^-$ (Guo and Cheng, 2018). Without the need for Gaussianity, Hastie et al. (2019) show that $\mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\} = \lim_{\lambda \to 0^+} \omega g(-\lambda) = \omega/(1-\omega)$. Then, a tiny small nonzero error term $1 - h_\beta^2$ can ruin the out-of-sample performance of OLS estimator (and more generally the ridge-less estimator) when ω is close to one. Ridge estimator $\widehat{\boldsymbol{\beta}}_R(\lambda)$ avoids the unboundedness problem of $\mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}$ by introducing a nonzero shrinkage term λ. In marginal estimator $\widehat{\boldsymbol{\beta}}_S$, the estimator of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is simply $\{\mathrm{Diag}(\boldsymbol{X}^T\boldsymbol{X})\}^{-1}$, which can be viewed as an extreme case of banded covariance estimator (Bickel and Levina, 2008) with zero bandwidth. Thus, $\widehat{\boldsymbol{\beta}}_S$ can avoid the issue of $\mathrm{tr}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\}$. However, the price is that $\widehat{\boldsymbol{\beta}}_S$ may have larger squared bias. See

Appendix B for more details, in which we illustrate the bias-variance decomposition using mean squared prediction errors of these estimators.

In summary, besides heritability, genetic correlation, and ω, the gap between true signal strength and $A_R^2(\lambda)$ is determined by the Stieltjes transform and its first order derivative. More importantly, the relative out-of-sample performance of these estimators highly depends on ω. When ω is large, all of them can become near-optimal for out-of-sample prediction. Therefore, choosing the optimal value of λ may become less important as dimension increases.

4.3.2 In-sample R-squared

In this section, we present the results for in-sample R^2 as a goodness-of-fit statistic, which is related to the performance of many in-sample applications of GWAS summary statistics (Barbeira et al., 2018). We find that in-sample R^2 has completely different patterns compared to out-of-sample R^2. The asymptotic results are summarized in the following theorem.

Theorem 4.3. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n, m_\beta, p) \to \infty$*, for any* $\omega \in (0, \infty)$*,* $h_\beta^2 \in (0, 1]$ *and* $\mathbf{\Sigma}$*, we have*

$$\begin{aligned} E_R^2(\lambda) &= E_B^2(\lambda/\omega) \\ &= \frac{\Big[h_\beta^2 \cdot \big\{1 - \lambda + \lambda^2 g(-\lambda)\big\} + (1 - h_\beta^2) \cdot \omega\big\{1 - \lambda g(-\lambda)\big\}\Big]^2}{h_\beta^2 \cdot \big\{1 - 2\lambda + 3\lambda^2 g(-\lambda) - \lambda^3 \dot{g}(-\lambda)\big\} + (1 - h_\beta^2) \cdot \omega\big\{1 - 2\lambda + \lambda^2 \dot{g}(-\lambda)\big\}} + o_p(1), \end{aligned}$$

where $\omega\{g(-\lambda) - \lambda^{-1}\} = v(-\lambda) - \lambda^{-1}$ *and* $\omega\{\dot{g}(-\lambda) - \lambda^{-2}\} = \dot{v}(-\lambda) - \lambda^{-2}$*. The* $E_R^2(\lambda)$ *is maximized as* $\lambda \to 0^+$*, and we have*

$$E_R^2(0^+) = \big\{h_\beta^2 + (1 - h_\beta^2) \cdot \frac{\omega + 1 - |\omega - 1|}{2}\big\} + o_p(1) = \begin{cases} E_O^2, & \text{if} \quad \omega < 1; \\ 1 + o_p(1), & \text{if} \quad \omega \geq 1, \end{cases}$$

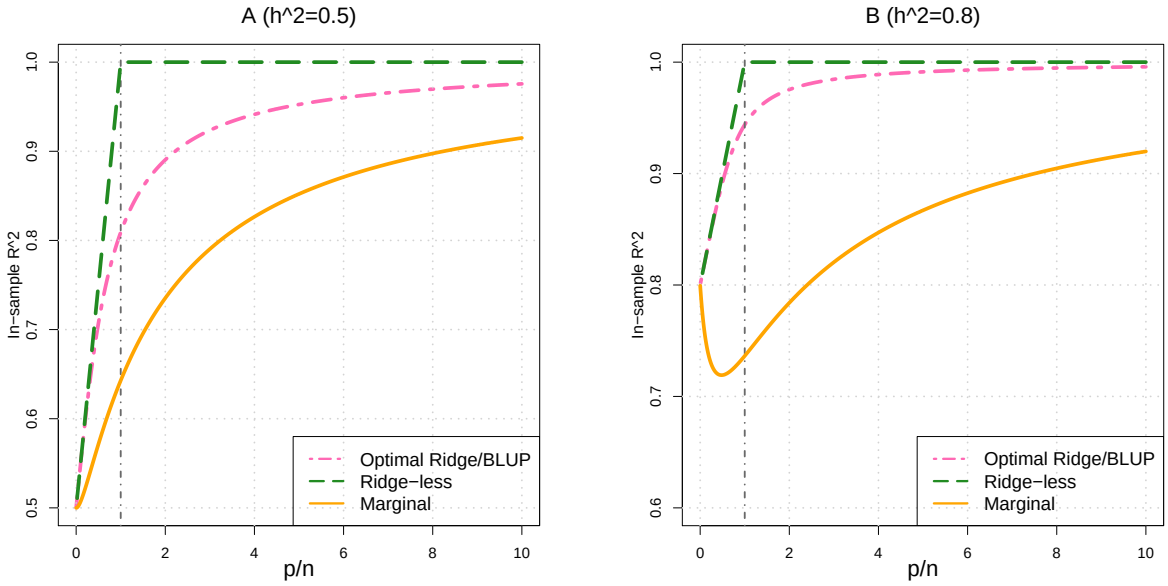

Figure 4.3: In-sample R-squared of ridge/BLUP-less estimators ($E_R^2(0^+) = E_B^2(0^+)$), optimal (out-of-sample) ridge/BLUP estimators ($E_R^2(\lambda^*) = E_B^2(\lambda^*/\omega)$), and marginal estimator (E_S^2) when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. We set $h_\beta^2 = (0.5, 0.8)$ in A and B, respectively.

where $E_O^2 = \left\{h_\beta^2 + (1 - h_\beta^2) \cdot \omega\right\} + o_p(1)$. *In addition, we have*

$$E_R^2(\lambda^*) = \frac{h_\beta^2}{1 - \lambda^* + \lambda^{*2} g(-\lambda^*)} + o_p(1).$$

When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, *the closed-form expression of* $E_R^2(\lambda^*)$ *is given by*

$$E_R^2(\lambda^*) = \frac{2h_\beta^6}{(1 - h_\beta^2) \cdot \left\{\sqrt{(\omega - h_\beta^2)^2 + 4\omega h_\beta^2(1 - h_\beta^2)} - \omega\right\} + h_\beta^2 \cdot (3h_\beta^2 - 1)} + o_p(1).$$

Theorem 4.3 has several implications. The optimal in-sample R-squared $E_R^2(0^+) \geq h_\beta^2$ for any $h_\beta^2 \in (0, 1]$ and is a linear function of $\omega \in (0, 1)$ (Figure 4.3). The term $(1 - h_\beta^2) \cdot \omega$ in E_O^2 represents the degree of model over-fitting due to spurious correlations (Fan et al., 2018). For $\omega > 1$, the limit of $E_R^2(0^+)$ is one, which indicates that the ridge-less estimator can have zero training error for $\boldsymbol{y}$ given any $h_\beta^2 \in (0, 1]$. On the other hand, E_S^2 may not be a monotone function of ω depending on h_β^2. When $h_\beta^2 \in (0, 0.5]$, E_S^2 increases with ω; when $h_\beta^2 \in (0.5, 1]$,

interestingly, E_S^2 decreases first as ω increases and can become much smaller than h_β^2.

4.4 Numerical experiments

4.4.1 Simulation

To numerically evaluate our asymptotic results, we first simulate data according to our modeling framework in Section 4.1 with $n = n_z = 2000$, and $\omega = 1.05$, 2, 4 and 8. Each entry of $\boldsymbol{X}$ and $\boldsymbol{Z}$ is a continuous variable that independently generated from $N(0, 1/p)$, and the ratio m/p is set to be 0.8. We simulate a trait with heritability 0.8 from model (4.1), and predict the same trait in the testing data (i.e., $h_\beta^2 = h_\eta^2 = 0.8$, $\varphi_{\beta\eta} = 1$, $\boldsymbol{\beta}_{(1)} = \boldsymbol{\eta}_{(1)}$). The nonzero genetic effects $\boldsymbol{\beta}_{(1)}$ and entries of $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_z$ are generated from Normal distribution according to Condition 4.2. We evaluate the following estimators:

(i) marginal estimator defined in model (2.2) (Marginal);

(ii) a meta-analyzed version of marginal estimator with weights equal to sample sizes (400 and 1600, respectively; Marginal-meta);

(iii) ridge estimator defined in model (2.3) with optimal regularizer λ^* (Ridge-Optimal);

(iv) ridge estimator with $n\lambda^*$ (Ridge-Over1);

(v) ridge estimator with $n^2\lambda^*$ (Ridge-Over2);

(vi) ridge estimator with λ^*/n (Ridge-Under1); and

(vii) ridge estimator with λ^*/n^2 (Ridge-Under2).

In addition, we examine three other methods in our settings including LASSO (Tibshirani, 1996), Elastic-Net (Zou and Hastie, 2005), and support vector machines (SVM) (Cortes and Vapnik, 1995). A total of 100 replicates is conducted, and we calculate the in-sample and out-of-sample R-squared (A^2 and E^2) defined in equation (4.5).

The results are summarized in Figure 4.4. As expected, the finite sample performance of marginal and ridge estimators supports our asymptotic results. For example, when $\omega = 1.05$, the optimal ridge estimator (Ridge-Optimal) clearly outperforms marginal estimator (Marginal), and marginal estimator has similar R^2 to ridge estimators with large λ (Ridge-

Over1/-Over2). On the other hand, ridge estimators with small λ (Ridge-Under1/-Under2) perform poorly for $\omega = 1.05$. These results indicate the importance of choosing the optimal λ when ω is close to one, and particularly shows that small λ should be avoided. However, when ω becomes 4, it is clear that marginal and all ridge estimators have similar R^2. In addition, meta-analyzed marginal estimator (Marginal-meta) shows no decay of prediction accuracy. The performance of LASSO and Elastic-Net is OK when ω is small, but becomes very poor as ω becomes large, suggesting that the methods designed with sparsity assumption should be used with caution in dense signal problems. SVM shows similar pattern to ridge estimators (Figure 4.4).

To mimic the linkage disequilibrium (LD) structure of SNP data, we also construct $\boldsymbol{\Sigma}$ with a block-diagonal structure (block size = 20). Features within the block have pair-wise correlation $\rho_b = 0.8$, and features belong to different blocks are independent. Other settings are exactly the same as in $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. The results are shown in Figures 4.6-4.7. Again, the performance of marginal and ridge estimators matches our theoretical limits, and the general pattern remains the same as in $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. In addition, the prediction accuracy is improved due to the feature-wise correlation, verifying that the decay of prediction accuracy due to dimensionality can be delayed by correlation among features.

4.4.2 UKB data simulation

Next, we perform simulation based on real SNP data from the UK Biobank (UKB) resources (Sudlow et al., 2015). There are $461,488$ common genotyped genetic variants (most of which are SNPs) after standard quality control (QC) procedures detailed in Appendix B. We randomly select $5,000$ or $10,000$ individuals of British ancestry as training samples, and test the prediction accuracy of these genetic variants on another $1,000$ randomly selected individuals of the same ancestry. Causal variants are randomly selected, and the number m is set to 470, 4700, $47,000$, and $235,000$, respectively. The nonzero genetic effects are independently generated from $N(0, 1/p)$, and we set $h^2_\beta = h^2_\eta = 0.8$, $\varphi_{\beta\eta} = 1$, and $\boldsymbol{\beta}_{(1)} = \boldsymbol{\eta}_{(1)}$. Marginal estimtaor is generated using PLINK (Purcell et al., 2007). Following practical

guidelines (Choi et al., 2018), we perform LD-based clumping for the marginal estimator via PLINK to obtain a list of relatively independent genetic variants for prediction. With the default window size (250 kb), we vary the clumping parameter C_r^2 and set it to 0.05, 0.1, 0.2, 0.3, 0.5, and 0.9. Smaller C_r^2 results in more stringent selection and more filtered variants by clumping. When $C_r^2 = 0.9$, most of the variants remain. The BLUP estimator is obtained from GCTA (Yang et al., 2011) using all genetic variants, in which the regularizer τ is estimated by the REML method (Jiang et al., 2016). The results are displayed in Figure 4.8. It is clear that the out-of-sample performance is stable across different signal sparsity m. When $n = 5,000$, the performance of BLUP and marginal estimators is in a similar range, though BLUP estimator is slightly better. After increasing the sample size to $10,000$, the difference between their performances becomes much more noticeable. This pattern suggests that the relative performance of the two estimators highly depends on ω. In addition, we find that results are not sensitive to the clumping parameter C_r^2 when genetic signals are dense.

4.5 Real data analysis

In this section, we present an imaging genetics example to predict brain subcortical structures in GWAS. We focus on 14 volumetric traits of seven left/right pairs of brain regions of interest (refer to as ROI volumes) including left/right thalamus proper, left/right caudate, left/right putamen, left/right pallidum, left/right hippocampus, left/right amygdala, and left/right accumbens area. The ROI volumes are quantified by magnetic resonance imaging (MRI) and are known to be associated with many brain-related cognitive and mental health traits (Miller et al., 2016). We use the UKB imaging samples as training data (Phase 1, $n = 9,868$; or Phases 1 and 2, $n = 19,629$) to predict the ROI volumes on subjects in three independent cohorts: Alzheimer's Disease Neuroimaging Initiative (ADNI) (Weiner et al., 2013), the Human Connectome Project (HCP) (Somerville et al., 2018), and Pediatric Imaging, Neurocognition, and Genetics (PING) (Jernigan et al., 2016). More details about data processing, quality control procedures, and cohort information can be found

in Appendix B. We perform prediction using both marginal (with $C_r^2 = 0.2$) and BLUP estimators, which are generated by PLINK and GCTA, respectively. The association between the predicted and observed phenotype is estimated in linear models, adjusting for the effects of age and sex. The associated partial R^2 is used to measure the prediction accuracy.

As shown in Figure 4.9, the performance of marginal and BLUP estimators is very similar when only the UKB Phase 1 data are used (Tables 4.1-4.3). However, after adding the UKB Phase 2 data, BLUP estimator perform better than marginal estimator for most of the traits (Tables 4.4-4.6). For instance, when $n = 9,868$, the median partial R^2 of BLUP and marginal estimators in the PING cohort are 0.80% and 0.69%, respectively. As n increases to $19,629$, the median partial R^2 become 1.44% and 0.78% for BLUP and marginal estimators, respectively. These results match our theoretical results that the relative out-of-sample performance of the marginal and BLUP estimators depends on ω. In practice, small partial R^2 are widely reported for brain-related traits (Bogdan et al., 2018), indicating that current GWAS sample size is still far from sufficient. Our analysis further suggests that, as we increase the GWAS sample size, BLUP estimator can benefit more and start to outperform marginal estimator. Thus, it might be better to estimate and share BLUP estimator instead of marginal estimator for neuroimaging traits prediction in the future when larger training GWAS is available.

4.6 Discussion

In GWAS, marginal screening has been a useful tool to improve our scientific understanding of complex traits. Recently, there is a pressing need to translate numerous scientific discoveries to clinical improvements, and one of the examples is to perform complex trait prediction directly using publicly available GWAS summary-level data. Motivated by these applications, we study high-dimensional dense signal prediction in large-scale GWAS context. Using the R-squared measure, we quantify the widespread gap between heritability and prediction accuracy for the popular GWAS marginal summary statistics and generalize the results to

Table 4.1: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in the ADNI cohort (n training$=9,868$).

ROI ID	R2-BLUP	R2-Marginal	P-value-BLUP	P-value-Marginal
left.thalamus.proper	7.389E-01	7.007E-01	1.523E-03	2.024E-03
left.caudate	2.232E-01	3.608E-01	1.209E-01	4.847E-02
left.putamen	1.981E+00	1.907E+00	1.984E-06	3.086E-06
left.pallidum	1.438E+00	1.109E+00	3.634E-05	2.917E-04
left.hippocampus	3.429E-01	2.724E-01	4.109E-02	6.871E-02
left.amygdala	5.338E-01	4.059E-01	9.355E-03	2.351E-02
left.accumbens.area	1.011E+00	1.198E+00	1.444E-04	3.491E-05
right.thalamus.proper	6.011E-01	6.502E-01	3.651E-03	2.496E-03
right.caudate	3.501E-01	3.571E-01	4.334E-02	4.131E-02
right.putamen	2.441E+00	2.618E+00	8.397E-08	2.812E-08
right.pallidum	1.642E+00	1.437E+00	1.366E-05	4.781E-05
right.hippocampus	3.000E-01	2.682E-01	5.874E-02	7.392E-02
right.amygdala	1.402E-01	1.986E-01	1.507E-01	8.708E-02
right.accumbens.area	1.293E+00	1.393E+00	8.150E-05	4.303E-05

cover cross-trait prediction and meta-analysis. We then examine and compare the class of ridge-type estimators, uncovering that such gap is a fundamental challenge for all these estimators and ω largely determines their relative performance. We also illustrate the different or even reverse behaviors of in-sample and out-of-sample R^2 in high-dimensions. Our theoretical results can be useful to evaluate the prediction accuracy in GWAS and other dense signal applications.

A few interesting future problems can be studied in the high-dimensional dense setting. First, ridge-type estimators represent linear shrinkage estimation on $\mathbf{\Sigma}$ and its inverse $\mathbf{\Sigma}^{-1}$. It might be interesting to explore whether we can improve the prediction accuracy with nonlinear shrinkage estimators, such as Ledoit and Wolf (2018). Second, if prior knowledge is known on the structure of $\mathbf{\Sigma}$, it is also possible to perform structured covariance estimation on $\mathbf{\Sigma}$, see Cai et al. (2016) for a review of this area. For example, since the $\mathbf{\Sigma}$ of SNP data is known to have a block-diagonal structure, it can be modeled as a bandable covariance matrix (Bickel and Levina, 2008) with fast decay of feature correlation as their physical distance

Table 4.2: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in the HCP cohort (n training$=9,868$).

ROI ID	R2-BLUP	R2-Marginal	P-value-BLUP	P-value-Marginal
left.thalamus.proper	2.517E-01	4.324E-01	5.571E-02	1.210E-02
left.caudate	4.154E-01	3.910E-01	2.582E-02	3.059E-02
left.putamen	5.630E-01	8.098E-01	5.563E-03	8.758E-04
left.pallidum	4.081E-01	5.314E-01	1.562E-02	5.780E-03
left.hippocampus	8.188E-01	4.570E-01	7.715E-04	1.209E-02
left.amygdala	1.701E-01	2.847E-01	1.165E-01	4.217E-02
left.accumbens.area	4.956E-01	9.989E-01	1.333E-02	4.332E-04
right.thalamus.proper	3.295E-01	4.245E-01	2.683E-02	1.195E-02
right.caudate	4.926E-01	5.277E-01	1.472E-02	1.157E-02
right.putamen	7.003E-01	9.672E-01	1.836E-03	2.477E-04
right.pallidum	6.297E-01	6.088E-01	3.340E-03	3.913E-03
right.hippocampus	6.783E-01	6.881E-01	2.311E-03	2.146E-03
right.amygdala	2.713E-02	2.656E-02	5.206E-01	5.251E-01
right.accumbens.area	1.287E-01	2.136E-01	2.184E-01	1.127E-01

increases. Indeed, as mentioned before, marginal estimator can be viewed as a special banded covariance estimator of $\mathbf{\Sigma}$ with zero bandwidth, which may represent an extreme estimator that over-bands $\mathbf{\Sigma}$. Finally, other extensions such as binary outcomes, time-to-event data, and SNP functional annotations and selections are also of great interest following the presented framework.

Table 4.3: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in the PING cohort (n training$=9,868$).

ROI ID	R2-BLUP	R2-Marginal	P-value-BLUP	P-value-Marginal
left.thalamus.proper	1.260E-01	4.284E-01	2.282E-01	2.612E-02
left.caudate	1.245E+00	1.161E+00	5.735E-04	8.818E-04
left.putamen	1.294E+00	1.593E+00	2.445E-04	4.645E-05
left.pallidum	9.033E-01	7.758E-01	1.663E-03	3.578E-03
left.hippocampus	2.459E+00	1.909E+00	2.820E-08	1.047E-06
left.amygdala	8.040E-02	3.024E-02	3.112E-01	5.346E-01
left.accumbens.area	5.018E-01	5.936E-01	2.362E-02	1.381E-02
right.thalamus.proper	1.703E-03	1.153E-01	8.882E-01	2.474E-01
right.caudate	1.481E+00	1.272E+00	1.735E-04	5.056E-04
right.putamen	1.057E+00	1.360E+00	9.127E-04	1.666E-04
right.pallidum	1.400E+00	1.213E+00	9.867E-05	2.917E-04
right.hippocampus	6.979E-01	4.280E-01	2.914E-03	1.987E-02
right.amygdala	5.709E-03	6.999E-04	7.895E-01	9.255E-01
right.accumbens.area	3.214E-01	3.139E-01	7.374E-02	7.722E-02

Table 4.4: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in the ADNI cohort (n training$=19,629$).

ROI ID	R2-BLUP	R2-Marginal	P-value-BLUP	P-value-Marginal
left.thalamus.proper	1.019E+00	1.084E+00	1.938E-04	1.212E-04
left.caudate	1.037E+00	7.481E-01	8.019E-04	4.451E-03
left.putamen	3.812E+00	2.433E+00	3.291E-11	1.314E-07
left.pallidum	2.118E+00	9.108E-01	5.086E-07	1.033E-03
left.hippocampus	5.549E-01	7.920E-03	9.316E-03	7.565E-01
left.amygdala	1.110E+00	4.695E-01	1.730E-04	1.482E-02
left.accumbens.area	1.314E+00	8.996E-01	1.440E-05	3.395E-04
right.thalamus.proper	8.978E-01	9.739E-01	3.759E-04	2.111E-04
right.caudate	5.633E-01	3.458E-01	1.033E-02	4.467E-02
right.putamen	4.401E+00	3.279E+00	4.519E-13	4.757E-10
right.pallidum	2.444E+00	1.307E+00	1.040E-07	1.056E-04
right.hippocampus	8.140E-01	1.410E-01	1.818E-03	1.952E-01
right.amygdala	3.438E-01	3.195E-01	2.430E-02	2.992E-02
right.accumbens.area	1.993E+00	1.558E+00	9.461E-07	1.498E-05

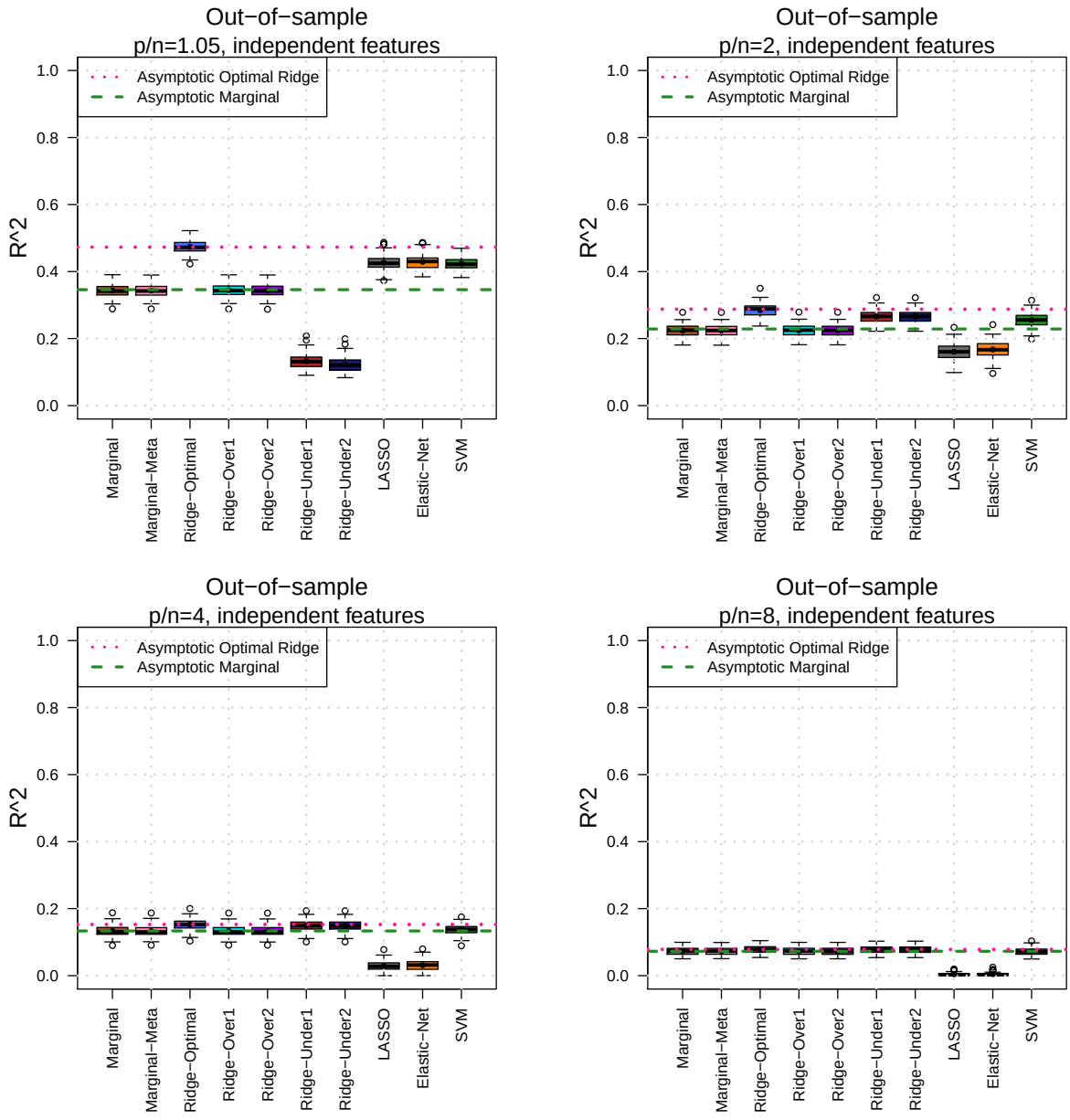

Figure 4.4: Out-of-sample R-squared of different estimators for independent features. Marginal: $\widehat{\boldsymbol{\beta}}_S$; Marginal-meta: meta-analyzed $\widehat{\boldsymbol{\beta}}_S$; Ridge-Optimal: $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$; Ridge-Over1: $\widehat{\boldsymbol{\beta}}_R(n\lambda^*)$; Ridge-Over2: $\widehat{\boldsymbol{\beta}}_R(n^2\lambda^*)$; Ridge-Under1: $\widehat{\boldsymbol{\beta}}_R(\lambda^*/n)$; Ridge-Under2: $\widehat{\boldsymbol{\beta}}_R(\lambda^*/n^2)$. We set $n = 2000$, and vary $\omega =$ from 1.05 to 8. The dash lines represent the asymptotic limits of ridge (red) and marginal (green) estimators.

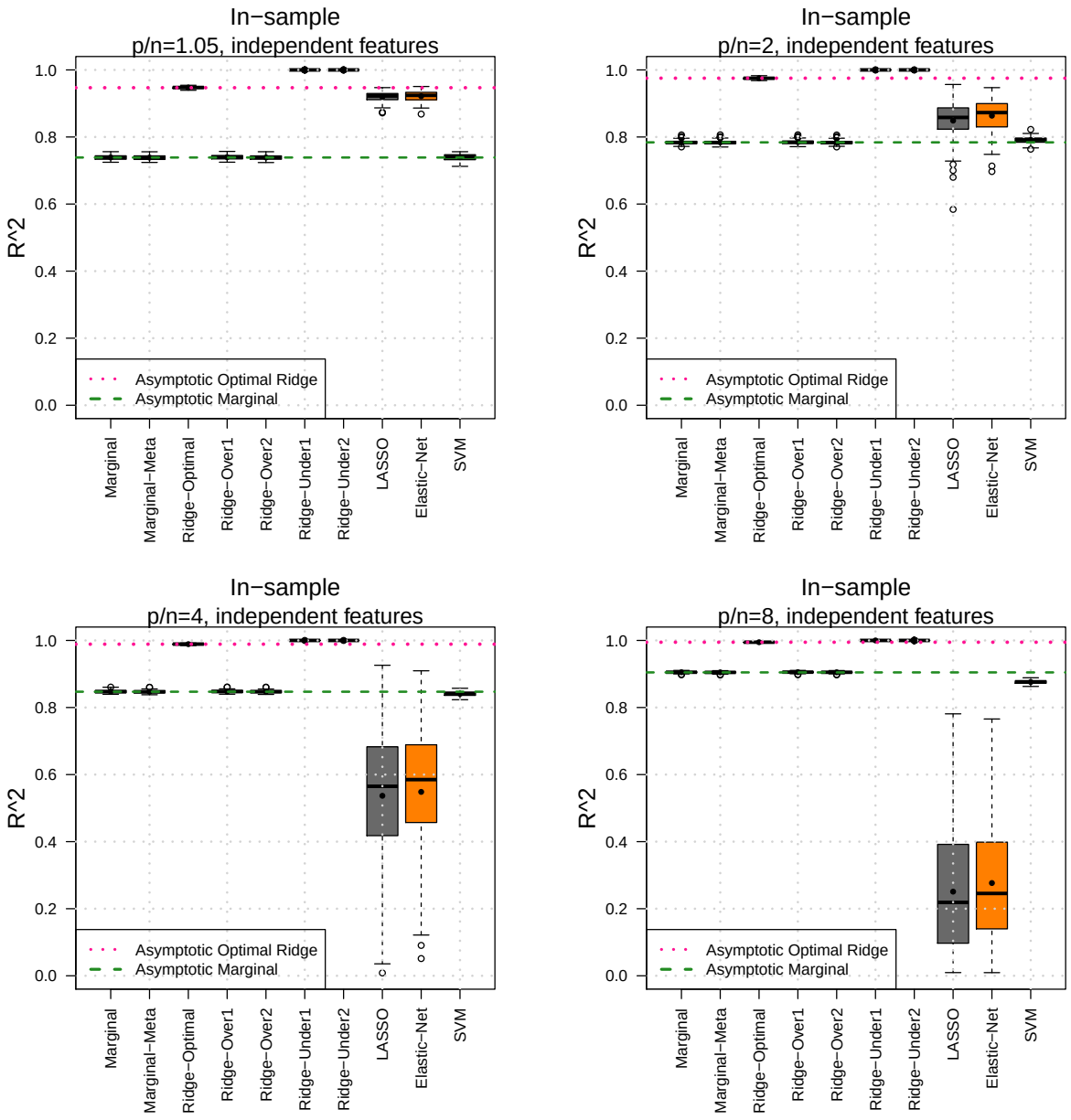

Figure 4.5: In-sample R-squared of different estimators for independent features. We set $n = 2000$, and vary $\omega =$ from 1.05 to 8. The dash lines represent the asymptotic limits of ridge (red) and marginal (green) estimators.

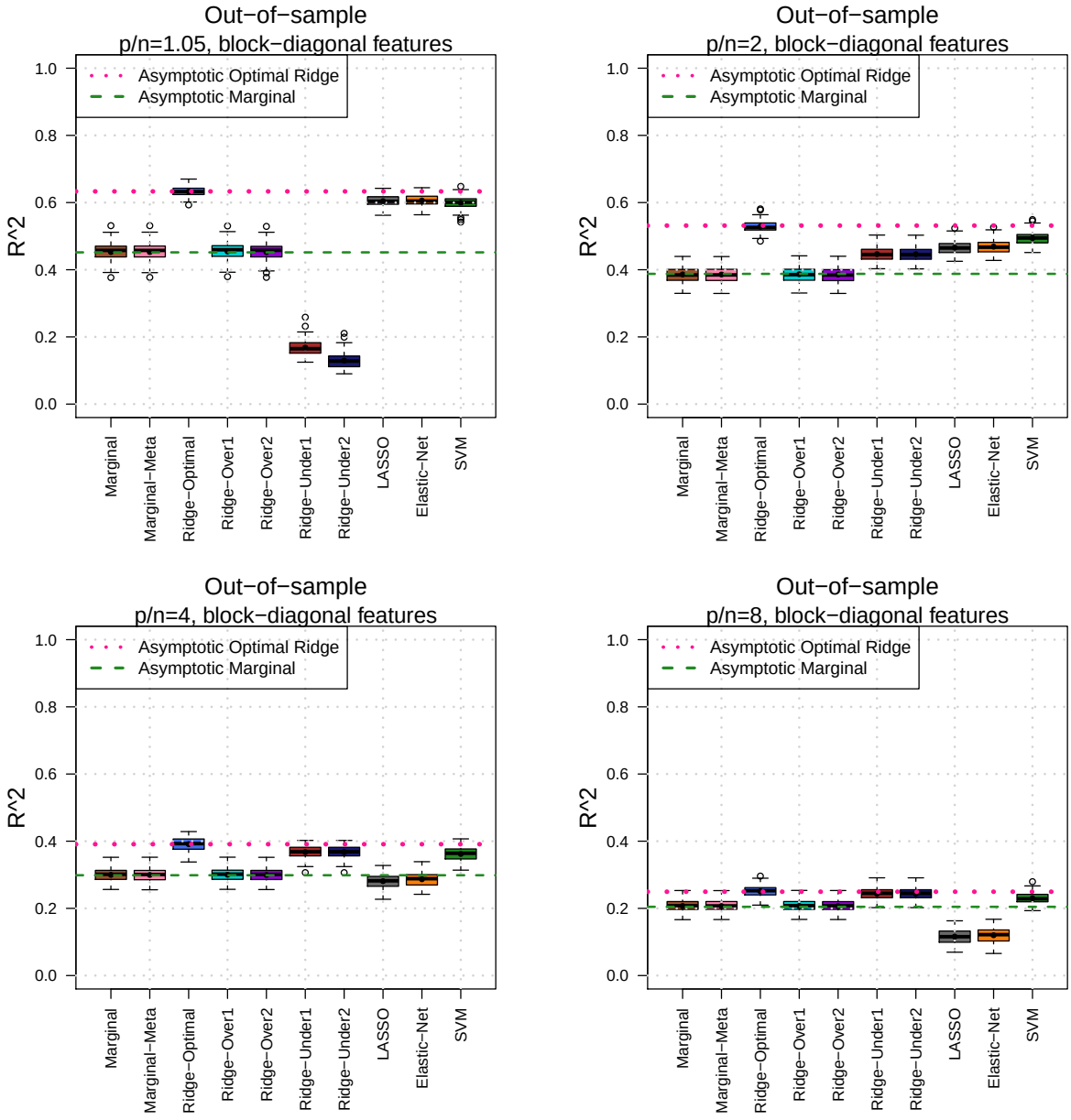

Figure 4.6: Out-of-sample R-squared of different estimators for features with block-diagonal correlation structure. We set $n = 2000$, and vary $\omega =$ from 1.05 to 8. The dash lines represent the asymptotic limits of ridge (red) and marginal (green) estimators.

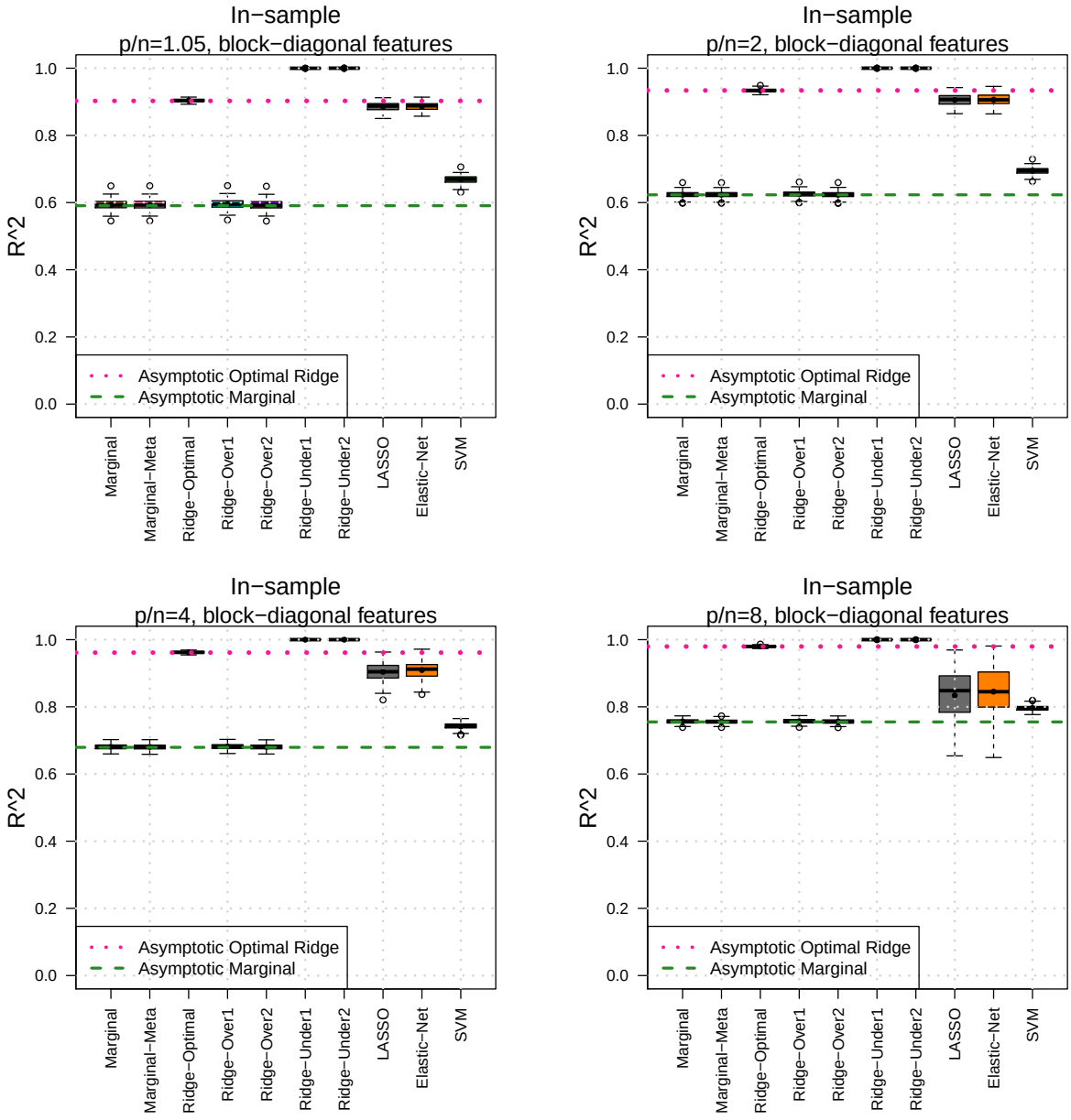

Figure 4.7: In-sample R-squared of different estimators for features with block-diagonal correlation structure. We set $n = 2000$, and vary ω from 1.05 to 8. The dash lines represent the asymptotic limits of ridge (red) and marginal (green) estimators.

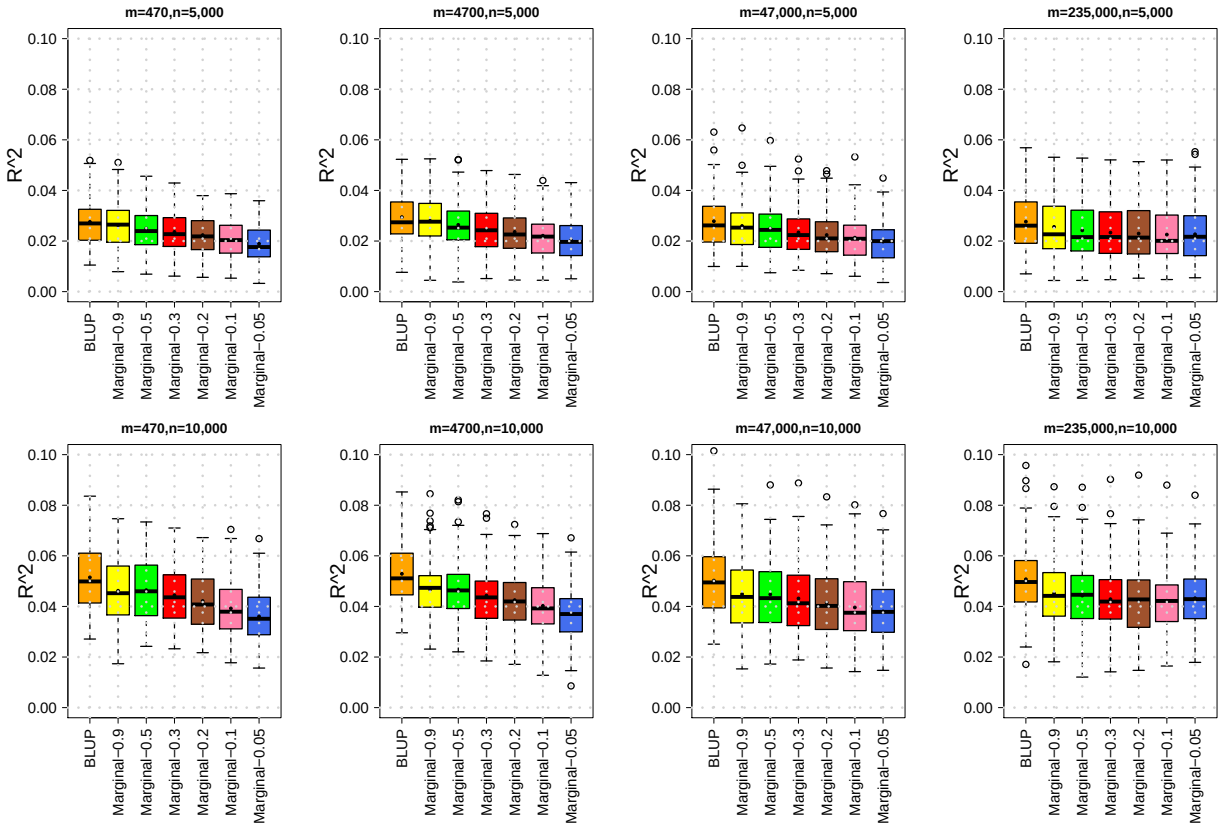

Figure 4.8: Out-of-sample R-squared of BLUP and margimal estimators across different sparsity m/p and sample size n. For marginal estimator, we try different clumping parameters: $0.05, 0.1, 0.2, 0.3, 0.4, 0.5$, and 0.9. We set $p = 461,499$, and vary ω from 470 to $235,000$. We set $n = 5,000$ in upper panels, and $n = 10,000$ in lower ones.

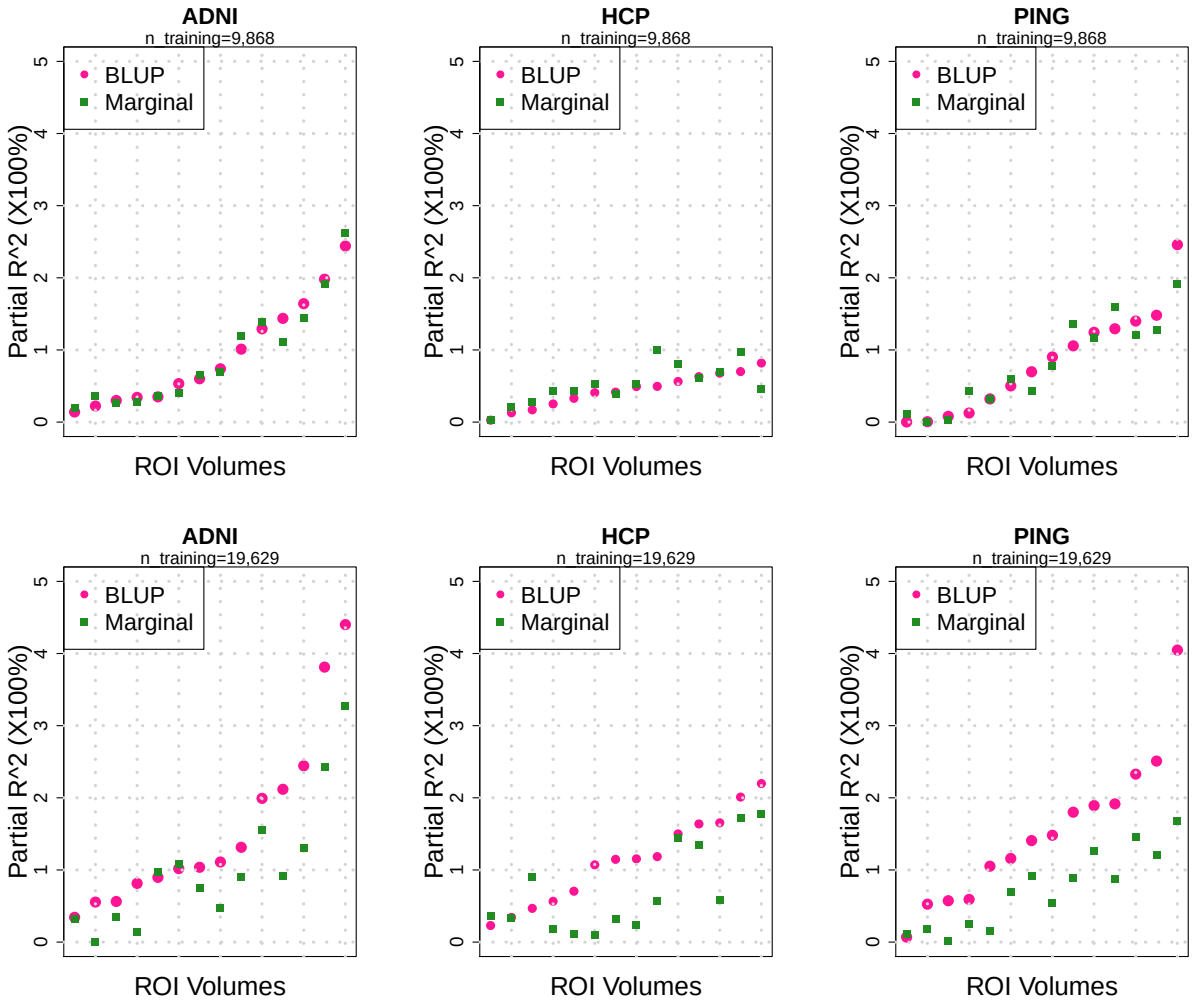

Figure 4.9: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in ADNI, HCP, and PING cohorts. BLUP: best linear unbiased prediction; Marginal: marginal estimator. Each point represents one ROI volume phenotype. The partial R-squared is estimated from linear regression while adjusting for the effects of age and gender. Upper panels: training with UKB Phase 1 data (n training=$9,868$); Lower panels: training with UKB Phases 1 and 2 data (n training=$19,629$).

Table 4.5: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in the HCP cohort (n training$=19,629$).

ROI ID	R2-BLUP	R2-Marginal	P-value-BLUP	P-value-Marginal
left.thalamus.proper	2.301E-01	3.635E-01	6.737E-02	2.144E-02
left.caudate	1.147E+00	3.185E-01	2.054E-04	5.103E-02
left.putamen	1.500E+00	1.440E+00	5.654E-06	8.749E-06
left.pallidum	5.658E-01	1.839E-01	4.396E-03	1.049E-01
left.hippocampus	2.010E+00	1.713E+00	1.219E-07	1.066E-06
left.amygdala	1.185E+00	5.638E-01	3.214E-05	4.211E-03
left.accumbens.area	1.653E+00	5.842E-01	5.726E-06	7.190E-03
right.thalamus.proper	4.680E-01	8.960E-01	8.292E-03	2.538E-04
right.caudate	7.054E-01	1.097E-01	3.490E-03	2.501E-01
right.putamen	2.198E+00	1.771E+00	2.909E-08	6.602E-07
right.pallidum	1.072E+00	1.050E-01	1.263E-04	2.316E-01
right.hippocampus	1.638E+00	1.342E+00	2.029E-06	1.738E-05
right.amygdala	3.421E-01	3.408E-01	2.239E-02	2.265E-02
right.accumbens.area	1.154E+00	2.400E-01	2.188E-04	9.272E-02

Table 4.6: Partial R-squared ($\times 100\%$) of 14 subcortical ROI volumes in the PING cohort (n training$=19,629$).

ROI ID	R2-BLUP	R2-Marginal	P-value-BLUP	P-value-Marginal
left.thalamus.proper	5.736E-01	1.514E-02	1.002E-02	6.762E-01
left.caudate	1.481E+00	5.364E-01	1.709E-04	2.405E-02
left.putamen	2.328E+00	1.463E+00	8.012E-07	9.537E-05
left.pallidum	1.916E+00	8.765E-01	4.341E-06	1.952E-03
left.hippocampus	4.048E+00	1.678E+00	7.854E-13	4.800E-06
left.amygdala	5.919E-01	2.444E-01	5.907E-03	7.727E-02
left.accumbens.area	1.159E+00	6.891E-01	5.680E-04	7.961E-03
right.thalamus.proper	5.260E-01	1.825E-01	1.336E-02	1.455E-01
right.caudate	1.407E+00	9.137E-01	2.536E-04	3.235E-03
right.putamen	1.892E+00	1.259E+00	8.688E-06	2.942E-04
right.pallidum	2.510E+00	1.205E+00	1.694E-07	3.057E-04
right.hippocampus	1.052E+00	1.486E-01	2.526E-04	1.705E-01
right.amygdala	6.951E-02	1.167E-01	3.515E-01	2.272E-01
right.accumbens.area	1.801E+00	8.941E-01	2.136E-05	2.815E-03

CHAPTER 5: ASSEMBLED RIDGE ESTIMATORS FOR GWAS DATA

5.1 General assembled ridge estimators

Assembled estimators can be efficiently generated by combining together the marginal estimator learned from training data and the feature covariance structure estimated on an independent reference panel. In this chapter, we extend our analysis on ridge-type estimators to investigate the prediction performance of assembled estimators. The assembled ridge estimator can be defined as

$$\widehat{\boldsymbol{\beta}}_A(\lambda) = \left(\boldsymbol{W}^T\boldsymbol{W} + \lambda n_w \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \propto \left(\boldsymbol{W}^T\boldsymbol{W} + \lambda n_w \boldsymbol{I}_p\right)^{-1}\widehat{\boldsymbol{\beta}}_S, \quad \lambda \in (0,\infty), \tag{5.1}$$

where $\boldsymbol{W}$ is a $n_w \times p$ SNP data matrix that is independent of $\boldsymbol{X}$. We consider two different versions of $\widehat{\boldsymbol{\beta}}_A(\lambda)$ that are common in practice: 1) $\boldsymbol{W} = \boldsymbol{Z}$ is from the testing data, donated as $\widehat{\boldsymbol{\beta}}_{A_1}(\lambda)$; and 2) $\boldsymbol{W}$ is independent of both training and testing data, donated as $\widehat{\boldsymbol{\beta}}_{A_2}(\lambda)$. Let $p/n_w \to \omega_w \in (0,\infty)$, we assume $\boldsymbol{W}$ also satisfies Condition 4.1 and have the following results about the out-of-sample R^2 of $\widehat{\boldsymbol{\beta}}_{A_1}(\lambda)$, donated as $A^2_{A_1}(\lambda)$.

Theorem 5.1. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n, n_z, n_w, p) \to \infty$*, for any* $\omega_w, \omega \in (0,\infty)$*,* $h^2_\beta, h^2_\eta \in (0,1]$*,* $\varphi_{\beta\eta} \in [-1,1]$*, and* $\boldsymbol{\Sigma}$*, we have*

$$A^2_{A_1}(\lambda) = h^2_\eta \varphi^2_{\beta\eta} \cdot \frac{\left[tr\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\}\right]^2 \cdot h^2_\beta}{tr\{(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^2\} \cdot h^2_\beta + \omega tr\{(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\}} + o_p(1),$$

where $\widehat{\boldsymbol{\Sigma}}_W = n_w^{-1} \cdot \boldsymbol{W}^T\boldsymbol{W}$.

For given traits, the prediction accuracy is determined by three traces $p^{-1}\text{tr}\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\}$, $\text{tr}\{(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^2\}$, and $\text{tr}\{(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\}$. We have

$$p^{-1}\text{tr}\{\boldsymbol{\Sigma}(\widehat{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\} \to_{a.s.} p^{-1}\text{tr}\{\boldsymbol{\Sigma}(a_w\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\},$$

where a_w is the unique positive solution of

$$1 - a_w = \omega\Big\{1 - \text{E}_{H(t)}\big(\frac{\lambda}{a_w t + \lambda}\big)\Big\}.$$

Clearly, $A^2_{A_1}(\lambda)$ is different from $A^2_R(\lambda)$, the asymptotic performance of the original ridge estimator $\widehat{\boldsymbol{\beta}}_R(\lambda)$. Now we consider the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have the following corollary.

Corollary 5.1. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n, n_z, n_w, p) \to \infty$*, for any* $\omega_w, \omega \in (0, \infty)$*,* $h^2_\beta, h^2_\eta \in (0, 1]$*,* $\varphi_{\beta\eta} \in [-1, 1]$*, when* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$*, we have*

$$A^2_{A_1}(\lambda) = h^2_\eta \varphi^2_{\beta\eta} \cdot \frac{h^2_\beta}{h^2_\beta + \omega} \cdot \frac{1}{1 - \dot{a}_w} + o_p(1),$$

where $\dot{a}_w = -(\omega_w a_w)/\{\omega_w \lambda + (a_w + \lambda)^2\}$.

We note that

$$A^2_S = h^2_\eta \varphi^2_{\beta\eta} \cdot \frac{h^2_\beta}{h^2_\beta + \omega} + o_p(1), \quad \text{and} \quad \dot{a}_w \leq 0,$$

it is clear that we have $A^2_{A_1}(\lambda) \leq A^2_S \leq A^2(\lambda)$ when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. The equality only holds for $\lambda \to \infty$ or $\omega_w = 0$, which is the classical low-dimensional case, or when the dimension of the dataset is extremely large. Thus, we have $A^2_{A_1}(\lambda) < A^2_S < A^2(\lambda)$ in general for $\lambda, \omega \in (0, \infty)$. In conclusion, assembling with external reference panel may not mimic the performance of original ridge estimator. In contrast, when the underlying $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, it can result in even lower prediction accuracy than marginal estimator. Next theorem gives the out-of-sample prediction R^2 of $\widehat{\boldsymbol{\beta}}_{A_2}(\lambda)$, donated as $A^2_{A_2}(\lambda)$.

Theorem 5.2. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ p) \to \infty$*, for any* $\omega_z, \omega \in (0,\infty)$*,* $h_\beta^2, h_\eta^2 \in (0,1]$*,* $\varphi_{\beta\eta} \in [-1,1]$*, and* $\boldsymbol{\Sigma}$*, we have*

$$A_{A_2}^2(\lambda) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\big\{1-\lambda \boldsymbol{Q}_{11}(a_z,\lambda)\big\}^2 \cdot h_\beta^2}{\{\boldsymbol{Q}_{12}(a_z,\lambda) - \lambda \boldsymbol{Q}_{22}(a_z,\lambda)\} \cdot h_\beta^2 + \omega\{\boldsymbol{Q}_{11}(a_z,\lambda) - \lambda \boldsymbol{Q}_{21}(a_z,\lambda)\}} + o_p(1),$$

where $\boldsymbol{Q}_{11}(a_z,\lambda) = p^{-1} tr\{(a_z\boldsymbol{\Sigma}+\lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\}$*,* $\boldsymbol{Q}_{12}(a_z,\lambda) = p^{-1} tr\{(a_z\boldsymbol{\Sigma}+\lambda \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^2\}$*,* $\boldsymbol{Q}_{21}(a_z,\lambda)$ $= p^{-1} tr\{(a_z\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_p)^{-2}(\boldsymbol{I}_p - \dot{a}_z\boldsymbol{\Sigma})\boldsymbol{\Sigma}\}$*, and* $\boldsymbol{Q}_{22}(a_z,\lambda) = p^{-1} tr\{(a_z\boldsymbol{\Sigma} + \lambda \boldsymbol{I}_p)^{-2}(\boldsymbol{I}_p - \dot{a}_z\boldsymbol{\Sigma})\boldsymbol{\Sigma}^2\}$*,* a_z *is the unique positive solution of*

$$1 - a_z = \omega_z\Big\{1 - \lambda \int_o^\infty (a_z t + \lambda)^{-1} dH(t)\Big\} = \omega_z\Big\{1 - E_{H(t)}\big(\frac{\lambda}{a_z t + \lambda}\big)\Big\},$$

and

$$\dot{a}_z = \frac{\omega_z E_{H(t)}\big\{\frac{a_z t}{(a_z t+\lambda)^2}\big\}}{-1 - \omega_z \lambda E_{H(t)}\big\{\frac{t}{(a_z t+\lambda)^2}\big\}}.$$

Moreover, we have

$$\boldsymbol{Q}_{11}(a_z,\lambda) = E_{H(t)}\big(\frac{t}{a_z t+\lambda}\big), \quad \boldsymbol{Q}_{12}(a_z,\lambda) = E_{H(t)}\big(\frac{t^2}{a_z t+\lambda}\big),$$
$$\boldsymbol{Q}_{21}(a_z,\lambda) = E_{H(t)}\big\{\frac{t(1-\dot{a}_z)}{(a_z t+\lambda)^2}\big\}, \quad \textit{and} \quad \boldsymbol{Q}_{22}(a_z,\lambda) = E_{H(t)}\big\{\frac{t^2(1-\dot{a}_z)}{(a_z t+\lambda)^2}\big\}.$$

For given traits, the prediction accuracy is determined by four traces $\boldsymbol{Q}_{11}$, $\boldsymbol{Q}_{12}$, $\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22}$. Consider the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have the following corollary.

Corollary 5.2. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ p) \to \infty$*, for any* $\omega_z, \omega \in (0,\infty)$*,* $h_\beta^2, h_\eta^2 \in (0,1]$*,* $\varphi_{\beta\eta} \in [-1,1]$*, when* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$*, we have*

$$A_{A_2}^2(\lambda) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{h_\beta^2}{h_\beta^2 + \omega} \cdot \frac{a_z^2 + \lambda}{a_z + \lambda} + o_p(1).$$

It can be shown that $a_z^2 \le a_z$. Thus, similar to $A_{A_1}^2(\lambda)$, we have $A_{A_2}^2(\lambda) < A_S^2$ in general for

$\lambda, \omega_z \in (0, \infty)$. We provide some insights into the assembling as follows. First, Theorems 5.1 and Theorems 5.2 show that the prediction accuracy of assembled ridge estimators is different from the original ridge estimators. As shown in later sections, $A^2_{A_1}(\lambda)$ and $A^2_{A_2}(\lambda)$ are very similar to each other, which of which are smaller than $A^2_R(\lambda)$. This finding suggests that assembling with testing dataset $\boldsymbol{Z}$ or external panel $\boldsymbol{W}$ have consistent performance. However, though assembling is very efficient, the price is that it can not achieve the performance of original estimators. Using the original ridge estimator directly from the original training dataset is always the best option. In the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we show that the assembled ridge estimators can even have worse performance than marginal estimators. Thus, assembling relatively independent features can result in worse performance and should be avoided. This motivates the block-wise assembling for block-diagonal covariance matrix, which is detailed in next section.

5.2 Block-wise assembling

In GWAS context, it is known that the covariance matrix of SNP data has a block-diagonal structure (Pasaniuc and Price, 2017). For example, the genome of the European ancestry can be divided into 1703 independent genomics regions, called LD blocks (Berisa and Pickrell, 2016). Therefore, assembling can be performed within each LD block. Suppose there are L blocks, the block-wise assembled ridge estimator in the l_{th} block can be defined as

$$\widehat{\boldsymbol{\beta}}_{Bl}(\lambda) = \left(\boldsymbol{W}_l^T \boldsymbol{W}_l + \lambda n_w \boldsymbol{I}_{p_l}\right)^{-1} \boldsymbol{X}_l^T \boldsymbol{y}, \quad \lambda \in (0, \infty), \tag{5.2}$$

for $l \in 1, \cdots, L$, where $\boldsymbol{W}_l$ and $\boldsymbol{X}_l$ are the $n_w \times p_l$ and $n \times p_l$ SNP data matrices for the l_{th} block, respectively, and p_l is the number of SNPs in the l_{th} block. Then the overall block-wise assembled ridge estimator is defined as $\widehat{\boldsymbol{\beta}}_B(\lambda) = (\widehat{\boldsymbol{\beta}}_{B1}(\lambda)^T, \cdots, \widehat{\boldsymbol{\beta}}_{BL}(\lambda)^T)^T$. Similar to $\widehat{\boldsymbol{\beta}}_A(\lambda)$, we let $\widehat{\boldsymbol{\beta}}_{B1}(\lambda)$ be the estimator when $\boldsymbol{W}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{Z}$, and $\widehat{\boldsymbol{\beta}}_{B2}(\lambda)$ be the estimator when $\boldsymbol{W} = \boldsymbol{Z}$.

Due to the efficiency, block-wise assembled ridge estimators similar to $\widehat{\boldsymbol{\beta}}_B(\lambda)$ have been

widely applied in complex trait prediction (Vilhjálmsson et al., 2015; Ge et al., 2019). When the block information is correct, block-wise assembling also have better performance than overall assembling, because it avoid assembling independent SNPs. Furthermore, here we propose to use the block-wise local principal components (BLPCs) instead of the raw SNP features in block-wise analysis. Specifically, for SNPs within the same block, we perform principal component (PC) analysis, and use the PCs for prediction. Intuitively, PC-based analysis has the following advantages: 1) SNPs within the same LD block can have very high correlation. PC-based analysis can avoid the potential collinearity issue in assembling; and 2) PCs can aggregate the small SNP effects together and reduce the feature dimension, which leads to improved performance. In the next section, we evaluate the performance of block-wise assembling estimators.

5.3 Numerical results

5.3.1 Asymptotic limits

To illustrate the finite sample performance of our asymptotic analysis, we simulate $n = 2,000$ individual samples in training data $\boldsymbol{X}$, external reference panel $\boldsymbol{W}$, and testing data $\boldsymbol{Z}$, respectively. We vary the number of SNPs p at $1,000$, $2,000$, $4,000$ or $8,000$ to reflect the different aspect ratio p/n. To mimic the LD structure of SNP data, we construct $\boldsymbol{\Sigma}$ with a block-diagonal structure (block size = 20). Features within the block have pair-wise correlation $\rho_b = 0.8$, and features belong to different blocks are independent. Each entry of $\boldsymbol{X}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ is generated from $N(0, 1)$. For training and testing data, the casual SNP effects $\boldsymbol{\beta}_{(1)} \sim MVN(\boldsymbol{0}, \boldsymbol{I}_m/p)$, and we set $m/p = 0.5$ and $h^2 = 0.6$ or $h^2 = 0.3$. The linear polygenic model (4.1) is used to generate $\boldsymbol{y}$ and $\boldsymbol{y}_z$. We illustrate the asymptotic results of the following estimators: 1) marginal estimator $\widehat{\boldsymbol{\beta}}_S$ (Marginal); 2) ridge estimator $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$, where λ^* is the optimal regularizer (Ridge); 3) assembled estimator $\widehat{\boldsymbol{\beta}}_{A2}(\lambda)$ with $\lambda = c \times \lambda^*$, where $c = 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4, 10^5$, respectively (Assembled-Z); 4) assembled estimator $\widehat{\boldsymbol{\beta}}_{A1}(\lambda)$ with $\lambda = c \times \lambda^*$ (Assembled-W); 5) block-wise estimator in original training

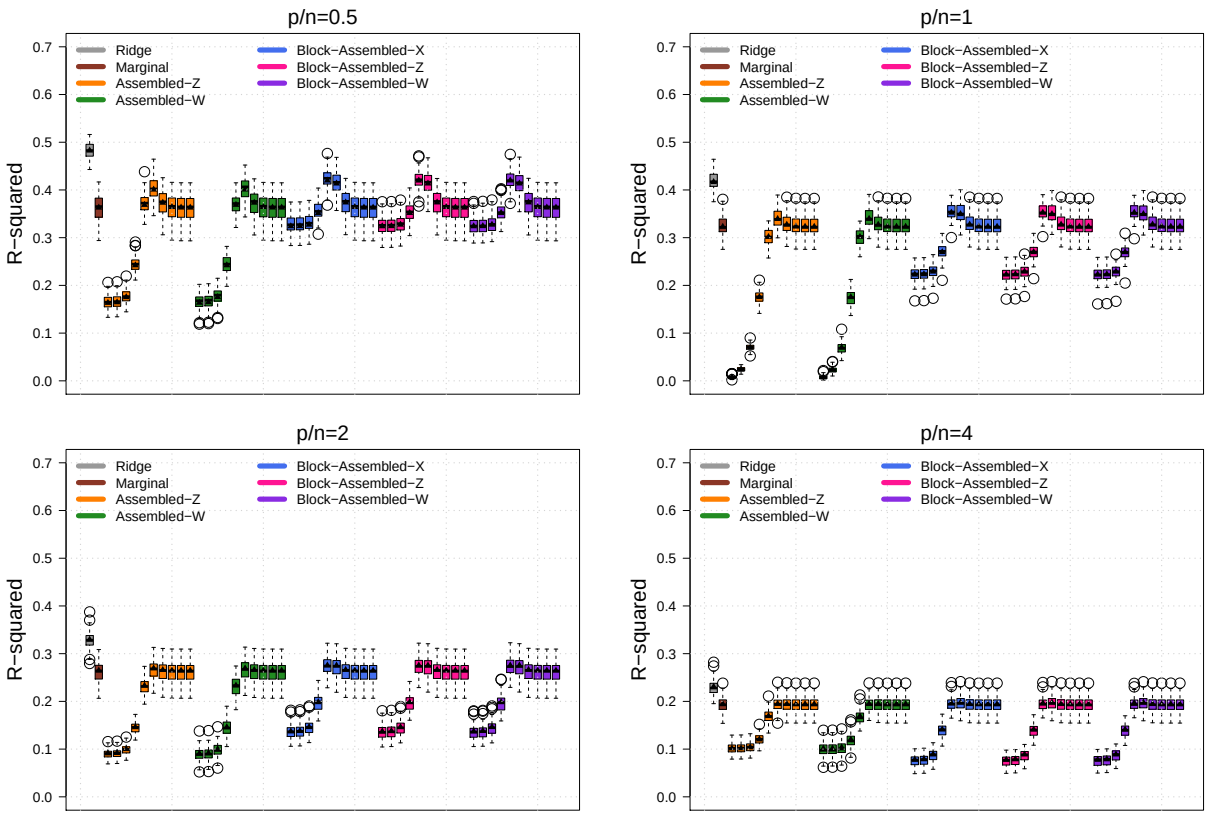

Figure 5.1: Out-of-sample R-squared of different estimators when $h^2 = 60\%$. Marginal: $\widehat{\boldsymbol{\beta}}_S$; Ridge: $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$; Assembled-Z: $\widehat{\boldsymbol{\beta}}_{A1}(\lambda)$; Assembled-W: $\widehat{\boldsymbol{\beta}}_{A2}(\lambda)$; Block-Assembled-X: $\widehat{\boldsymbol{\beta}}_{B0}(\lambda)$; Block-Assembled-Z: $\widehat{\boldsymbol{\beta}}_{B1}(\lambda)$; Block-Assembled-W: $\widehat{\boldsymbol{\beta}}_{B2}(\lambda)$. We set $n = 2000$, and $p/n = 0.5$, 1, 2, and 4, respectively. For the last five estimator, we try a series of $\lambda = c \times \lambda^*$, with $c = 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4$, and 10^5, respectively. The dots are asymptotic performances according to our theoretical results.

data $\boldsymbol{X}$ with $\lambda = c \times \lambda^*$, donated as $\widehat{\boldsymbol{\beta}}_{B0}(\lambda)$ (Block-Assembled-X); 6) block-wise assembled estimator $\widehat{\boldsymbol{\beta}}_{B2}(\lambda)$ with $\lambda = c \times \lambda^*$ (Assembled-Z); and 7) block-wise assembled estimator $\widehat{\boldsymbol{\beta}}_{B1}(\lambda)$ with $\lambda = c \times \lambda^*$ (Assembled-W). A total of 100 replications are conducted for each simulation condition.

Figure 5.1 displays the performance of these estimators across a series of p/n ratios when $h^2 = 60\%$. As expected, the performances are close to their corresponding asymptotic results for all estimators. The performance of $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$ is better than the one of $\widehat{\boldsymbol{\beta}}_{A1}(\lambda)$ and $\widehat{\boldsymbol{\beta}}_{A2}(\lambda)$, regardless of the choice of λ. This suggests that assembled estimators, though efficient, are sub-optimal compared to the original ridge estimator directly from the training data.

Moreover, when a small λ is chosen, assembled estimators can even perform worse than the marginal estimator. This is particularly true when p/n is close to one, in which case over-fitting with small regularizer could be dangerous and have very bad performance (Guo and Cheng, 2018; Hastie et al., 2019).

In contrast, block-wise estimators $\widehat{\boldsymbol{\beta}}_{B0}(\lambda)$, $\widehat{\boldsymbol{\beta}}_{B1}(\lambda)$, and $\widehat{\boldsymbol{\beta}}_{B2}(\lambda)$ have very similar performance. Therefore, in block-wise analysis, assembled estimators $\widehat{\boldsymbol{\beta}}_{B1}(\lambda)$ and $\widehat{\boldsymbol{\beta}}_{B2}(\lambda)$ can be not only extremely efficient, but have near the same performance as the original estimator $\widehat{\boldsymbol{\beta}}_{B0}(\lambda)$. Interestingly, Figure 5.1 also shows that $\widehat{\boldsymbol{\beta}}_{B0}(\lambda)$ has lower prediction accuracy than $\widehat{\boldsymbol{\beta}}_{R}(\lambda)$. Therefore, even we know the $\boldsymbol{\Sigma}$ has a block-diagonal structure and the block location information is correct, the block-wise estimator of $\boldsymbol{\Sigma}$ has lower prediction accuracy than the sample covariance estimator which ignores the structure of $\boldsymbol{\Sigma}$. On the other hand, $\widehat{\boldsymbol{\beta}}_{B}(\lambda)$ performs better than $\widehat{\boldsymbol{\beta}}_{A}(\lambda)$, especially when p/n is relatively small. This suggests that block-wise assembling is efficient and also leads to better prediction than overall assembling. The performance for the case with $h^2 = 30\%$ is presented in Figure 5.2, which shows similar patterns.

5.3.2 UK biobank data simulation

We perform additional simulation with real SNP data from the UK Biobank (UKB) resources (Sudlow et al., 2015; Bycroft et al., 2018). We download the imputed genotype data and apply the following quality control (QC) procedure: excluding subjects with more than 10% missing genotypes, only including SNPs with MAF > 0.01, genotyping rate $> 90\%$, and passing Hardy-Weinberg test (p-value $> 1 \times 10^{-7}$). The QC'ed data contains $8,932,279$ SNPs over $488,371$ subjects, from which we randomly select $110,000$ unrelated individuals of British ancestry to perform the simulation. Among the $110,000$ selected samples, $100,000$ are randomly picked and used as training samples, and prediction performance is evaluated with the remaining $10,000$ testing individuals. Moreover, we constrain our analysis to $653,122$ SNPs that are overlapped with the SNPs in the HapMap3 reference panel (Consortium et al., 2010), which is a popular choice in assembling analysis to balance the prediction accuracy and

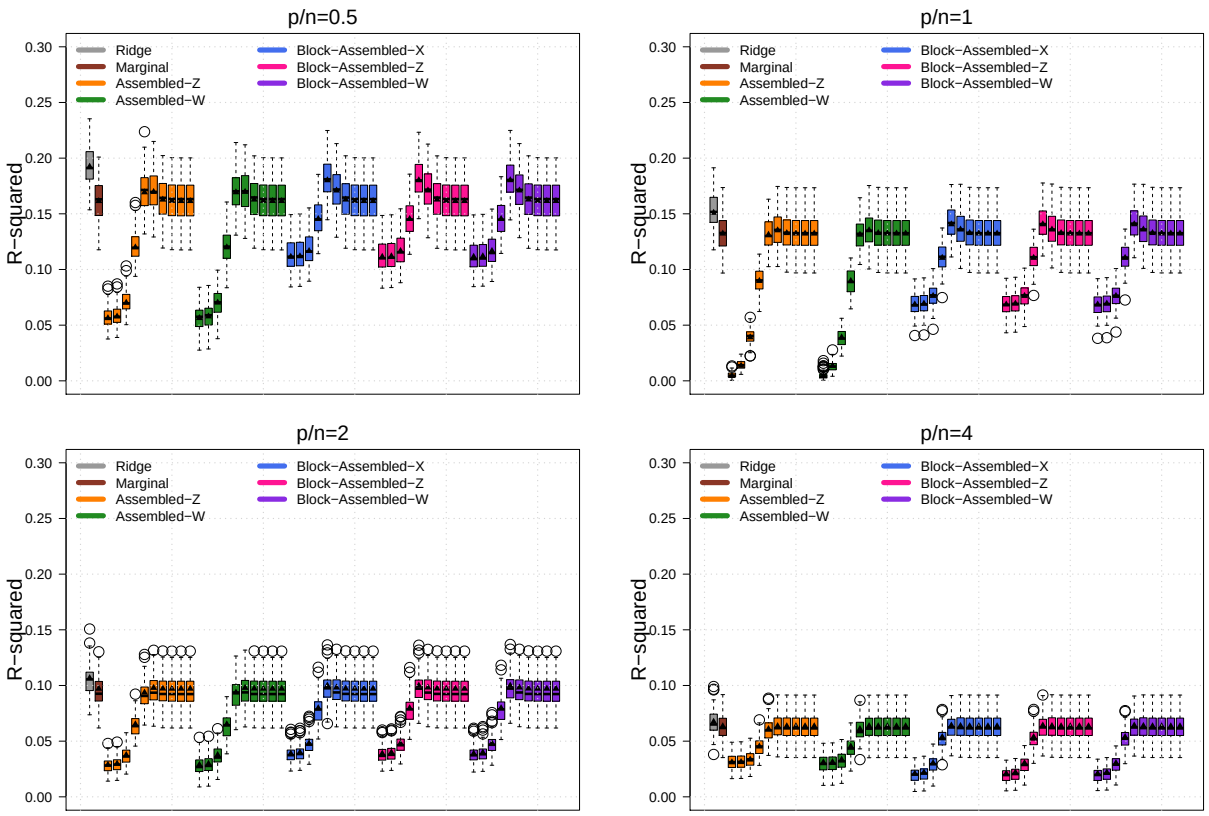

Figure 5.2: Out-of-sample R-squared of different estimators when $h^2 = 30\%$. Marginal: $\widehat{\boldsymbol{\beta}}_S$; Ridge: $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$; Assembled-Z: $\widehat{\boldsymbol{\beta}}_{A1}(\lambda)$; Assembled-W: $\widehat{\boldsymbol{\beta}}_{A2}(\lambda)$; Block-Assembled-X: $\widehat{\boldsymbol{\beta}}_{B0}(\lambda)$; Block-Assembled-Z: $\widehat{\boldsymbol{\beta}}_{B1}(\lambda)$; Block-Assembled-W: $\widehat{\boldsymbol{\beta}}_{B2}(\lambda)$. We set $n = 2000$, and $p/n = 0.5$, 1, 2, and 4, respectively. For the last five estimator, we try a series of $\lambda = c \times \lambda^*$, with $c = 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4$, and 10^5, respectively. The dots are asymptotic performances according to our theoretical results.

computational burden (Ge et al., 2019). We randomly select 200,000 SNPs to be causal SNPs. The nonzero SNP effects are independently generated from $N(0, 1)$, and the heritability h^2 is set to 60%.

We first perform the following SNP-based methods: 1) SNP marginal screening using the fastGWA toolset (Jiang et al., 2019) (SNP-Marginal); 2) SNP marginal screening for relatively independent pruned SNPs only (250 kb, pruning r-squared 0.3) (SNP-Marginal-Prune); 3) SNP block-wise assembled ridge estimator using the PCS toolset (Ge et al., 2019) (SNP-Ref-Ridge). Next, we perform block-wise local PCA within each of the 1,703 independent genomics regions (Berisa and Pickrell, 2016). For each block, we keep the top BLPCs that

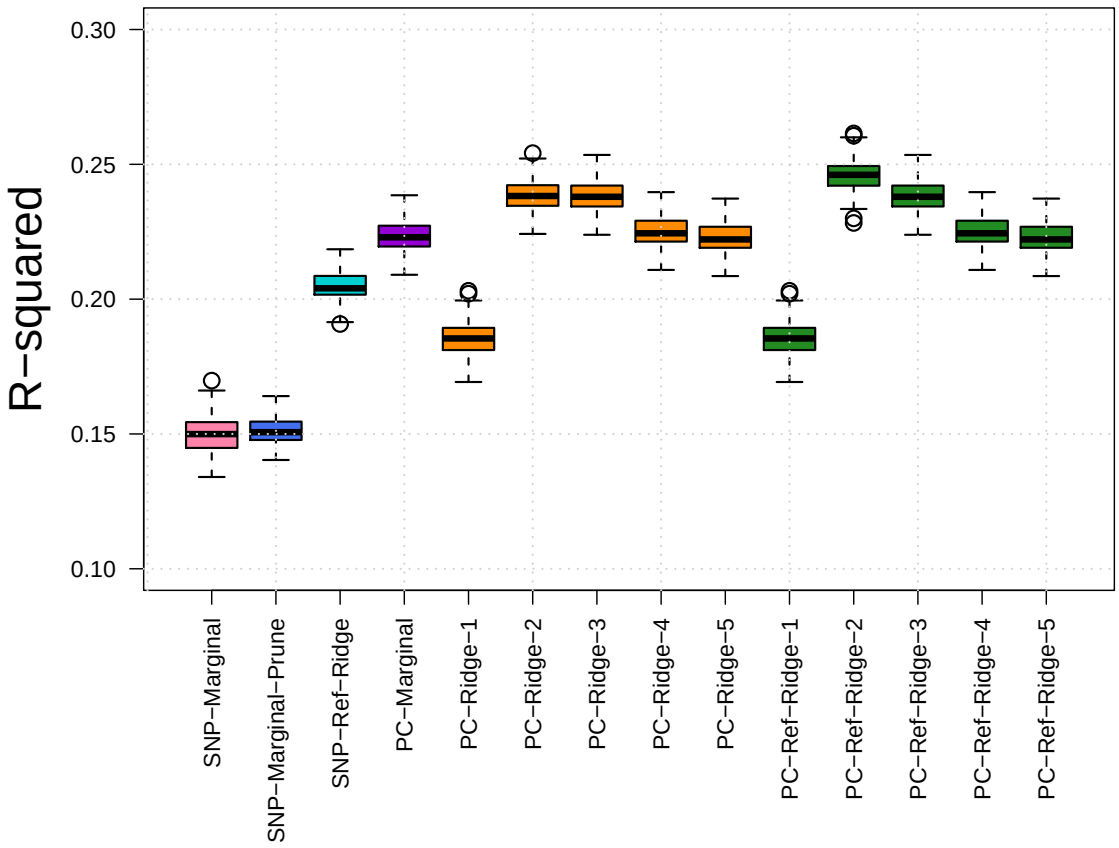

Figure 5.3: Out-of-sample R-squared of different estimators in UK biobank data simulation. SNP-Marginal: SNP marginal screening; SNP-Marginal-Prune: screening for pruned SNPs; SNP-Ref-Ridge: block-wise SNP assembled ridge estimator; PC-Marginal: PC marginal screening; PC-Ridge: block-wise PC ridge estimator (with varying regularizers); and PC-Ref-Ridge: block-wise PC assembled ridge estimator (with varying regularizers).

can cumulatively explain more than 50% genetic variation. In totally, we have $58,768$ BLPCs to be used in downstream prediction. We consider the following BLPC-based methods: 1) PC marginal screening (PC-Marginal); 2) PC block-wise ridge estimator using training data, with $\lambda = c \times \lambda^*$, $\lambda^* = 58,768/100,000$, $c = 10, 1, 0.1, 0.01$, and 0, respectively (PC-Ridge); and 3) PC block-wise assembled ridge estimator using testing data, with $\lambda = c \times \lambda^*$, $\lambda^* = 58,768/10,000$, $c = 10, 1, 0.1, 0.01$, and 0, respectively (PC-Ref-Ridge).

The results are displayed in Figure 5.3. First, PC-Marginal has better performance than

all SNP-based methods, suggesting the advantage of using PCs instead of the SNP features in complex traits prediction. Second, both PC-Ridge and PC-Ref-Ridge can further improve the prediction accuracy on top of PC-Marginal and the two have similar performance. The optimal regularizer of ridge estimator (λ^*), or a slightly smaller one than it, works well in our simulations. In conclusion, BLPC-based estimators can outperform SNP-based estimators in out-of-sample prediction. In addition, assembled ridge estimator can have similar performance to the original block-wise ridge estimator.

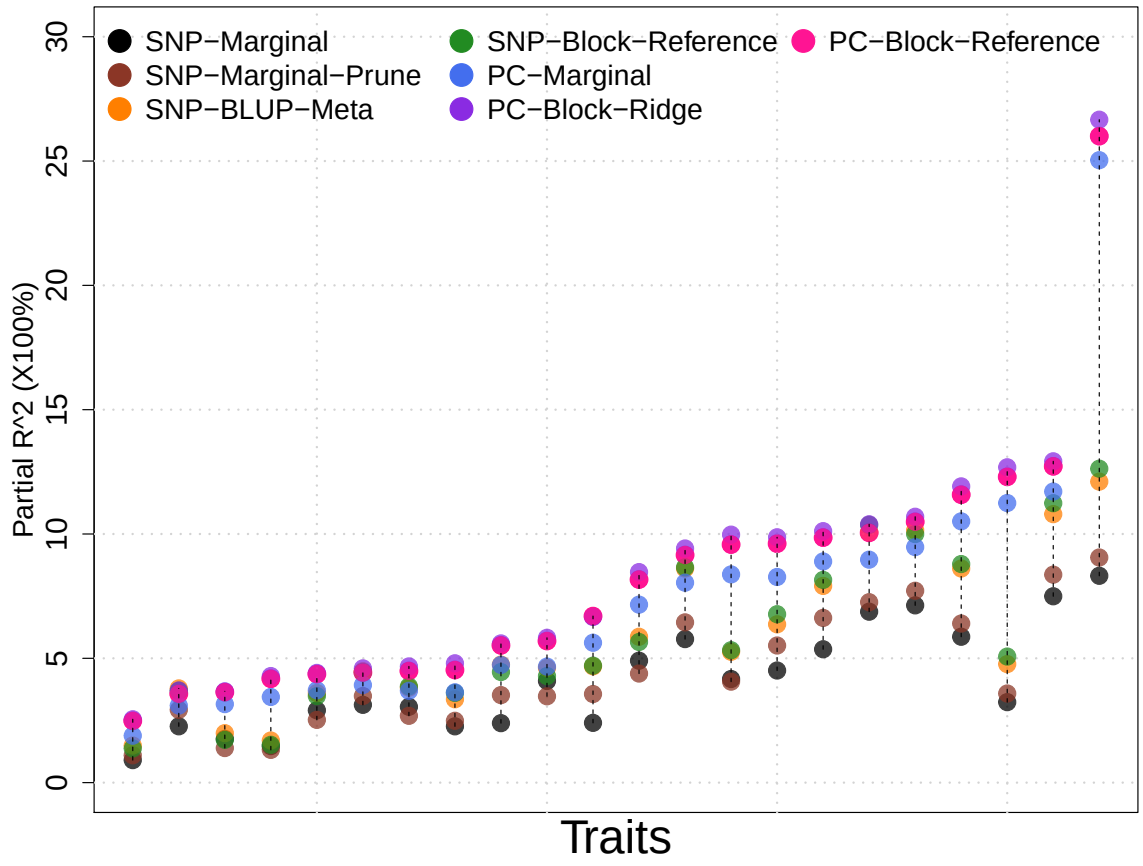

Figure 5.4: Out-of-sample R-squared of different estimators for UK Biobank complex traits. SNP-Marginal: SNP marginal screening; SNP-Marginal-Prune: marginal screening for pruned SNPs; SNP-BLUP-Meta: SNP BLUP estimator; SNP-Ref-Ridge: SNP block-wise assembled ridge estimator; PC-Marginal: PC marginal screening; PC-Block-Ridge: PC block-wise ridge estimator; and PC-Block-Reference: PC block-wise assembled ridge estimator.

5.4 Real data examples

We perform prediction on 22 complex UKB traits from several different trait domains, including anthropometric traits, blood traits, and cardiorespiratory traits (Table 5.1). Similar to the above simulation analysis, we focus on the unrelated individuals of British ancestry, and randomly select $10,000$ individuals as testing subjects, and the remaining subjects are used for training. The mean sample size of the training data for these traits is $300,000$, see Table 5.1 for details. The prediction accuracy on the testing samples is measured by the (partial) R-squared of PRS from the linear model. For each block, we keep the top BLPCs that can cumulatively explain more than 65% genetic variation, resulting in $101,507$ BLPCs in total. We adjust the same set of covariates in the training and testing data, including age, age-squared, sex, and the top 40 genetic PCs provided by UKB (Bycroft et al., 2018).

Similar to the simulation analysis, we evaluate and compare the following methods 1) SNP marginal screening using the fastGWA toolset (Jiang et al., 2019) (SNP-Marginal); 2) SNP marginal screening for relatively independent SNPs after SNP pruning (250 kb, pruning r-squared 0.3) (SNP-Marginal-Prune); 3) SNP block-wise assembled ridge estimator using the PCS toolset (Ge et al., 2019) (SNP-Block-Reference); 4) PC marginal screening (PC-Marginal); 5) PC block-wise ridge estimator, with $\lambda = 101,507/n$ (PC-Block-Ridge); and 6) PC block-wise assembled ridge estimator using testing data, with $\lambda = 0.01 \times 101,507/10,000$ (PC-Block-Reference). In addition, we also compare with the BLUP estimator from the GCTA tooset (Yang et al., 2011). BLUP estimator is very computationally intensive and require hugh CPU memory given such a sample size. We reduce the computational burden by a ten-fold meta-analysis (SNP-BLUP-Meta).

Figure 5.4 displays the out-of-sample R^2 of different methods on these complex traits. It is clear that BLPC-based methods have better performance on most of these complex traits. For example, the prediction accuracy of height can be improved from 12.63% to 25.03% by PC-Marginal, and to 26.66% by PC-Block-Ridge (Tables 5.2 and 5.3). Moreover, the performance of PC-Block-Ridge and PC-Block-Reference are very similar. In conclusion, the

pattern of a wide variety of complex human traits matches well with the simulation results. We find that most of these complex traits can be better predicted using BLPCs and external reference panel can work very similar to the original training data. Such information can be valuable in real genetics prediction of these complex traits.

Table 5.1: Information of the UK Biobank complex traits.

Trait	Training sample size	Data-Field ID
Heel_bone_mineral_density	198,817	78
Hand_grip_strength_left	344,429	46
Hand_grip_strength_right	344,466	47
Waist_circumference	345,190	48
Hip_circumference	345,143	49
Height	345,023	50
BMI	344,700	21001
Basal_metabolic_rate	339,913	23105
Weight	344,818	21002
Body_fat_percentage	339,734	23099
Hair_color	345,769	1747
Balding_pattern	159,610	2395
LDL	329,976	30780
HDL	303,383	30760
Eosinophill_count	335,648	30150
Platelet_count	336,241	30080
Platelet_distribution_width	336,074	30110
Red_blood_cell_count	336,244	30010
Red_blood_cell_distribution_width	336,244	30070
White_blood_cell_count	336,240	30000
Diastolic_blood_pressure	325,845	4079
Systolic_blood_pressure	325,844	4080
Pulse_rate	325,845	102
FEV	259,993	20150
FVC	259,993	20151
Vascular_heart_problems	345,765	6150
Fluid_intelligence_score	138,094	20016
Digits_remembered	53,999	4282
Time_to_identify_matches	343,719	20023
Depression_sum_score	111,517	138
Neuroticism_sum_score	276,175	100060
Ever_smoked	344,582	20160
Smoking_status	344,539	20116

Table 5.2: Prediction accuracy of SNP-based estimators on UK Biobankcomplex traits.

Trait	Marginal	Marginal-Prune	BLUP-Meta	Block-Reference
Hand_grip_strength_left	1.47	1.32	1.69	1.52
Hand_grip_strength_right	1.76	1.39	1.99	1.72
Waist_circumference	4.90	4.39	5.87	5.66
Hip_circumference	5.77	6.45	8.61	8.66
Height	8.32	9.05	12.11	12.63
BMI	6.88	7.26	10.07	10.37
Basal_metabolic_rate	3.24	3.58	4.77	5.07
Weight	5.87	6.40	8.61	8.79
Body_fat_percentage	4.17	4.06	5.26	5.32
HDL	5.36	6.62	7.91	8.15
Eosinophill_count	2.39	3.52	4.75	4.45
Platelet_count	7.50	8.36	10.80	11.24
Platelet_distribution_width	7.13	7.72	10.15	10.00
Red_blood_cell_count	4.51	5.51	6.37	6.77
White_blood_cell_count	2.40	3.56	4.68	4.70
Diastolic_blood_pressure	2.26	2.49	3.35	3.61
Systolic_blood_pressure	3.05	2.69	3.82	3.87
Pulse_rate	4.09	3.47	4.61	4.29
FEV	2.26	2.93	3.79	3.69
FVC	3.13	3.48	4.48	4.44
PEF	0.90	1.10	1.49	1.38
Vascular_heart_problems	2.91	2.52	3.60	3.48

Table 5.3: Prediction accuracy of BLPC-based estimators on UK Biobankcomplex traits.

Trait	PC-Marginal	PC-Block-Ridge	PC-Block-Reference
Hand_grip_strength_left	3.45	4.29	4.18
Hand_grip_strength_right	3.16	3.67	3.64
Waist_circumference	7.16	8.47	8.17
Hip_circumference	8.04	9.42	9.15
Height	25.03	26.66	26.00
BMI	8.97	10.39	10.05
Basal_metabolic_rate	11.25	12.68	12.30
Weight	10.51	11.91	11.58
Body_fat_percentage	8.37	9.97	9.58
HDL	8.90	10.11	9.86
Eosinophill_count	4.73	5.60	5.52
Platelet_count	11.71	12.92	12.72
Platelet_distribution_width	9.47	10.69	10.49
Red_blood_cell_count	8.27	9.87	9.61
White_blood_cell_count	5.63	6.66	6.70
Diastolic_blood_pressure	3.63	4.80	4.53
Systolic_blood_pressure	3.69	4.67	4.49
Pulse_rate	4.67	5.82	5.70
FEV	3.10	3.71	3.57
FVC	3.92	4.60	4.43
PEF	1.89	2.55	2.48
Vascular_heart_problems	3.72	4.41	4.36

APPENDIX A: TECHNICAL DETAILS OF CHAPTER 3

Main proofs

In this section, we highlight the key steps and results to prove our main theorems.

Proposition A.1. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, if $m_{\alpha\eta}, m_\alpha$, and $m_\eta \to \infty$ as $min(n_1, n_3, p) \to \infty$, then we have*

$$\frac{\big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T\big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)}{n_3 m_\eta \cdot \sigma_\eta^2/p + n_3 \cdot \sigma_{\epsilon_\eta}^2} = 1 + o_p(1),$$
$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)}{\{n_1 n_3 m_\alpha(p - m_\alpha) + n_1 n_3 m_\alpha(m_\alpha + n_1)\} \cdot \sigma_\alpha^2/p + n_1 n_3 p \cdot \sigma_{\epsilon_\alpha}^2} = 1 + o_p(1).$$

Further if $p/(n_1 n_3) \to 0$, then we have

$$\frac{\big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)}{n_1 n_3 m_{\alpha\eta} \cdot \sigma_{\alpha\eta}/p} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$G_{\alpha\eta} = \sqrt{\frac{n_1}{n_1 + p/h_\alpha^2}} \cdot h_\eta \cdot \varphi_{\alpha\eta} + o_p(1).$$

Then Theorem 3.1 holds for $a \in (0, 1)$. When $a \in [1, \infty]$, i.e., $p/(n_1 n_3) \not\to 0$, we note

$$\big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) = O_p\big\{(n_1^{1/2} n_3^{1/2} m_{\alpha\eta} p^{1/2} + n_1 n_3 m_{\alpha\eta}) \cdot \sigma_{\alpha\eta}/p\big\}.$$

It follows that

$$\begin{aligned} G_{\alpha\eta}^2 &= O_p\left[\frac{(n_1 n_3 m_{\alpha\eta}^2 p + n_1^2 n_3^2 m_{\alpha\eta}^2) \cdot \sigma_{\alpha\eta}^2/p^2}{(n_3 m_\eta \cdot \sigma_\eta^2/p + n_3 \cdot \sigma_{\epsilon_\eta}^2) \cdot \big\{n_1 n_3 m_\alpha(n_1 + p) \cdot \sigma_\alpha^2/p + n_1 n_3 p \cdot \sigma_{\epsilon_\alpha}^2\big\}}\right] \\ &= O_p\left\{\frac{n_1 n_3 p + n_1^2 n_3^2}{n_3^2 n_1(n_1 + p/h_\alpha^2)/h_\eta^2} \cdot \varphi_{\alpha\eta}^2\right\} = O_p(\frac{1}{n_3}). \end{aligned}$$

Thus, Theorem 3.1 is proved.

Proposition A.2. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, if* $m_{\alpha\beta}, m_\alpha$, *and* $m_\beta \to \infty$ *as* $min(n_1, n_2, n_3, p) \to \infty$, *then we have*

$$\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{\{n_2n_3m_\beta(p-m_\beta) + n_2n_3m_\beta(m_\beta+n_2)\}\cdot\sigma_\beta^2/p + n_2n_3p\cdot\sigma_{\epsilon_\beta}^2} = 1 + o_p(1).$$

Further if $p^2/(n_1n_2n_3) \to 0$, *then we have*

$$\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)}{n_1n_2n_3m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p} = 1 + o_p(1).$$

It follows that Theorem 3.2 holds for $a \in (0,1)$. When $a \in [1,\infty]$, we have

$$\begin{aligned}
G_{\alpha\beta}^2 &= O_p\bigg[\frac{(n_1n_2n_3m_{\alpha\beta}^2p^2 + n_1^2n_2^2n_3^2m_{\alpha\beta}^2)\cdot\sigma_{\alpha\beta}^2/p^2}{\big\{n_2n_3m_\beta(n_2+p)\cdot\sigma_\beta^2/p + n_2n_3p\cdot\sigma_{\epsilon_\beta}^2\big\}\cdot\big\{n_1n_3m_\alpha(n_1+p)\cdot\sigma_\alpha^2/p + n_1n_3p\cdot\sigma_{\epsilon_\alpha}^2\big\}}\bigg] \\
&= O_p\bigg\{\frac{n_1n_2n_3p^2 + n_1^2n_2^2n_3^2}{n_3^2n_1n_2(n_1+p/h_\alpha^2)(n_2+p/h_\beta^2)}\bigg\} = O_p\bigg\{\frac{p^2}{n_3(n_1+p/h_\alpha^2)(n_2+p/h_\beta^2)}\bigg\} \\
&= O_p\bigg\{\frac{p^2}{n_3(n_1+p)(n_2+p)}\bigg\}.
\end{aligned}$$

Thus, Theorem 3.2 is proved.

Proposition A.3. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, if* $m_{\alpha\beta}, m_\alpha$, *and* $m_\beta \to \infty$ *as* $min(n_1, n_2, p) \to \infty$, *then we have*

$$\begin{aligned}
&\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)}{\{n_1m_\alpha(n_1+m_\alpha) + n_1m_\alpha(p-m_\alpha)\}\cdot\sigma_\alpha^2/p + n_1p\cdot\sigma_{\epsilon_\alpha}^2} = 1 + o_p(1), \\
&\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{\big\{n_2m_\beta(n_2+m_\beta) + n_2m_\beta(p-m_\beta)\big\}\cdot\sigma_\beta^2/p + n_2p\cdot\sigma_{\epsilon_\beta}^2} = 1 + o_p(1).
\end{aligned}$$

Further if $p/(n_1n_2) \to 0$, *then we have*

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{n_1n_2m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p} = 1 + o_p(1).$$

Therefore Theorem 3.3 holds for $a \in (0,1)$. When $a \in [1,\infty]$, we have

$$\widehat{\varphi}^2_{\alpha\beta} = O_p\Bigg[\frac{(n_1n_2m^2_{\alpha\beta}p + n_1^2n_2^2m^2_{\alpha\beta})\cdot\sigma^2_{\alpha\beta}/p^2}{\big\{n_2m_\beta(n_2+p)\cdot\sigma^2_\beta/p + n_2p\cdot\sigma^2_{\epsilon_\beta}\big\}\cdot\big\{n_1m_\alpha(n_1+p)\cdot\sigma^2_\alpha/p + n_1p\cdot\sigma^2_{\epsilon_\alpha}\big\}}\Bigg]$$
$$= O_p\Bigg\{\frac{n_1n_2p + n_1^2n_2^2}{n_1n_2(n_1+p/h^2_\alpha)(n_2+p/h^2_\beta)}\Bigg\} = O_p\Bigg\{\frac{p}{(n_1+p)(n_2+p)}\Bigg\}.$$

Thus, Theorem 3.3 is proved. Corollary 3.2 follows from the proposition below.

Proposition A.4. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, if* $min(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$, $min(q_{\alpha\eta}, q_{\alpha1}, q_{\alpha2}) \to \infty$ *when* $min(n_1, n_3, p) \to \infty$, *then we have*

$$\frac{V_{T\alpha}}{\{n_1n_3m_\alpha q_{\alpha2} + n_1n_3q_{\alpha1}(m_\alpha+n_1)\}\cdot\sigma^2_\alpha/p + n_1n_3q_\alpha\cdot\sigma^2_{\epsilon_\alpha}} = 1 + o_p(1).$$

Further if $\{m^2_{\alpha\eta}(q_{\alpha1}+q_{\alpha2})\}/(q^2_{\alpha\eta}n_1n_3) \to 0$, *then we have*

$$\frac{C_{T\alpha\eta}}{n_1n_3q_{\alpha\eta}\cdot\sigma_{\alpha\eta}/p} = 1 + o_p(1).$$

More SNP screening results

In this section, we provide more analysis for SNP screening. Given a threshold $c_\beta > 0$, let $q_\beta = p\cdot\pi_\beta = q_{\beta1} + q_{\beta2}$ ($\pi_\beta \in (0,1]$) be the number of top-ranked SNPs selected for $\boldsymbol{y}_\beta$, among which there are $q_{\beta1}$ true causal SNPs and the remaining $q_{\beta2}$ are null SNPs, and we let $q_{\alpha\beta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$. Thus, $\min(q_{\beta1}, q_{\alpha1}) \geq q_{\alpha\beta}$. The SNP data are defined accordingly. We write $\boldsymbol{Z}_{(1)} = [\boldsymbol{Z}_{(11)}, \boldsymbol{Z}_{(12)}]$, $\boldsymbol{Z}_{(2)} = [\boldsymbol{Z}_{(21)}, \boldsymbol{Z}_{(22)}]$, $\boldsymbol{W}_{(1,\beta)} = [\boldsymbol{W}_{(11,\beta)}, \boldsymbol{W}_{(12,\beta)}]$, and $\boldsymbol{W}_{(2,\beta)} = [\boldsymbol{W}_{(21,\beta)}, \boldsymbol{W}_{(22,\beta)}]$. Here $\boldsymbol{Z}_{(11)}$ and $\boldsymbol{W}_{(11,\beta)}$ are the selected $q_{\beta1}$ causal SNPs of $\boldsymbol{y}_\beta$, and $\boldsymbol{Z}_{(21)}$ and $\boldsymbol{W}_{(21,\beta)}$ are the selected $q_{\beta2}$ null SNPs of $\boldsymbol{y}_\beta$. In addition, we let $\widehat{\boldsymbol{\beta}}_{(1)} = [\widehat{\boldsymbol{\beta}}_{(11)}, \widehat{\boldsymbol{\beta}}_{(12)}]$, and $\widehat{\boldsymbol{\beta}}_{(2)} = [\widehat{\boldsymbol{\beta}}_{(21)}, \widehat{\boldsymbol{\beta}}_{(22)}]$, where $\widehat{\boldsymbol{\beta}}_{(11)}$ corresponds to the selected causal SNPs of $\boldsymbol{y}_\beta$, and $\widehat{\boldsymbol{\beta}}_{(21)}$ corresponds to the selected null ones. Similar to

$G_{T\alpha\eta}$, we have

$$G_{T\alpha\beta} = \frac{\big(\boldsymbol{W}_{(11,\beta)}\widehat{\boldsymbol{\beta}}_{(11)} + \boldsymbol{W}_{(21,\beta)}\widehat{\boldsymbol{\beta}}_{(21)}\big)^T\big(\boldsymbol{W}_{(11,\alpha)}\widehat{\boldsymbol{\alpha}}_{(11)} + \boldsymbol{W}_{(21,\alpha)}\widehat{\boldsymbol{\alpha}}_{(21)}\big)}{\|\boldsymbol{W}_{(11,\beta)}\widehat{\boldsymbol{\beta}}_{(11)} + \boldsymbol{W}_{(21,\beta)}\widehat{\boldsymbol{\beta}}_{(21)}\| \cdot \|\boldsymbol{W}_{(11,\alpha)}\widehat{\boldsymbol{\alpha}}_{(11)} + \boldsymbol{W}_{(21,\alpha)}\widehat{\boldsymbol{\alpha}}_{(21)}\|} = \frac{C_{T\alpha\beta}}{V_{T\alpha} \cdot V_{T\beta}},$$

where

$$C_{T\alpha\beta} = \big\{\boldsymbol{W}_{(11,\beta)}\boldsymbol{Z}_{(11)}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big) + \boldsymbol{W}_{(21,\beta)}\boldsymbol{Z}_{(21)}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)\big\}^T$$
$$\big\{\boldsymbol{W}_{(11,\alpha)}\boldsymbol{X}_{(11)}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) + \boldsymbol{W}_{(21,\alpha)}\boldsymbol{X}_{(21)}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\big\}$$

and $V_{T\beta} = \big\|\boldsymbol{W}_{(11,\beta)}\boldsymbol{Z}_{(11)}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big) + \boldsymbol{W}_{(21,\beta)}\boldsymbol{Z}_{(21)}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)\big\|$.

Proposition A.5. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, if min*$(m_{\alpha\beta}, m_\alpha, m_\beta) \to \infty$, *min*$(q_{\alpha\beta}, q_{\alpha1}, q_{\alpha2}, q_{\beta1}, q_{\beta2}) \to \infty$ *when min*$(n_1, n_2, n_3, p) \to \infty$, *then we have*

$$\frac{V_{T\beta}}{\{n_2n_3m_\beta q_{\beta2} + n_2n_3q_{\beta1}(m_\beta + n_2)\} \cdot \sigma_\beta^2/p + n_1n_3q_\beta \cdot \sigma_{\epsilon_\beta}^2} = 1 + o_p(1).$$

Further if $\{m_{\alpha\beta}^2(q_{\alpha1} + q_{\alpha2})(q_{\beta1} + q_{\beta2})\}/(q_{\alpha\beta}^2 n_1n_2n_3) \to 0$, *then we have*

$$\frac{C_{T\alpha\beta}}{n_1n_2n_3q_{\alpha\beta} \cdot \sigma_{\alpha\beta}/p} = 1 + o_p(1).$$

Following PropositionA.5, we have the consistency result for $G_{T\alpha\beta}$.

Corollary A.1. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose that min*$(m_{\alpha\beta}, m_\alpha, m_\beta) \to \infty$ *and min*$(q_{\alpha\beta}, q_{\alpha1}, q_{\alpha2}, q_{\beta1}, q_{\beta2}) \to \infty$ *as min*$(n_1, n_2, n_3, p) \to \infty$. *Further if* $\big\{m_{\alpha\beta}^2(q_{\alpha1} + q_{\alpha2})(q_{\beta1} + q_{\beta2})\big\}/(q_{\alpha\beta}^2 n_1n_2n_3) \to 0$, *then we have*

$$G_{T\alpha\beta} = \varphi_{\alpha\beta} + \left(\sqrt{\frac{n_1m_\alpha}{n_1q_{\alpha1} + m_\alpha q_\alpha/h_\alpha^2} \cdot \frac{n_2m_\beta}{n_2q_{\beta1} + m_\beta q_\beta/h_\beta^2}} \cdot \frac{q_{\alpha\beta}}{m_{\alpha\beta}} - 1\right) \cdot \varphi_{\alpha\beta} + o_p(1).$$

Corollary A.1 shows the trade-off of SNP screening for $G_{T\alpha\beta}$. Given n_1, n_2, m_α, m_β, $m_{\alpha\beta}$, h_α, and h_β, the potential bias of $G_{T\alpha\beta}$ is affected by q_α, $q_{\alpha1}$, q_β, $q_{\beta1}$ and $q_{\alpha\beta}$. As more

SNPs are selected, the numerator of $q_{\alpha\beta}/m_{\alpha\beta}$ increases with $q_{\alpha\beta}$, while the denominator of $\sqrt{(n_1 m_\alpha)/(n_1 q_{\alpha 1} + m_\alpha q_\alpha/h_\alpha^2) \cdot (n_2 m_\beta)/(n_2 q_{\beta 1} + m_\beta q_\beta/h_\beta^2)}$ increases with $\sqrt{q_\alpha}$ and $\sqrt{q_\beta}$ (also $\sqrt{q_{\alpha 1}}$ and $\sqrt{q_{\beta 1}}$). In the optimistic case where $q_{\alpha\beta} = m_{\alpha\beta}$, $q_\alpha = q_{\alpha 1} = m_\alpha$ and $q_\beta = q_{\beta 1} = m_\beta$, $G_{T\alpha\beta}$ reduces to

$$\sqrt{\frac{n_1}{n_1 + m_\alpha/h_\alpha^2} \cdot \frac{n_2}{n_2 + m_\beta/h_\beta^2}} \cdot \varphi_{\alpha\beta},$$

which is the theoretical upper limit. On the other hand, suppose $q_{\alpha\beta}/q_{\alpha 1} \approx m_{\alpha\beta}/m_\alpha$ and $q_{\alpha\beta}/q_{\beta 1} \approx m_{\alpha\beta}/m_\beta$, when $n_1 = o(m_\alpha)$, $n_2 = o(m_\beta)$, i.e., the causal SNPs and null SNPs are totally mixed, we have $q_{\alpha 1}/q_\alpha \approx m_\alpha/p$, $q_{\beta 1}/q_\beta \approx m_\beta/p$, and

$$G_{T\alpha\beta} \approx \sqrt{\frac{n_1}{n_1 p + p^2/h_\alpha^2} \cdot \frac{n_2}{n_2 p + p^2/h_\alpha^2} \cdot q_\alpha q_\beta} \cdot \varphi_{\alpha\beta},$$

which increases with q_α and q_β. Therefore, as $q_\alpha = q_\beta = p$, $G_{T\alpha\beta}$ reaches its upper bound

$$\sqrt{\frac{n_1}{n_1 + p/h_\alpha^2} \cdot \frac{n_2}{n_2 + p/h_\beta^2}} \cdot \varphi_{\alpha\beta}.$$

More proofs

In this section, we provide more details for overlapping samples.

Proposition A.6. *Under polygenic model (3.4) and Conditions 3.1 - 3.3, suppose $m_{\alpha\eta}, m_\alpha$, and $m_\eta \to \infty$ as $(n_1 + n_s), (n_3 + n_s), p \to \infty$, then we have*

$$\frac{\boldsymbol{y}_\eta^T \boldsymbol{y}_\eta}{(n_3 + n_s) m_\eta \cdot \sigma_\eta^2/p + (n_3 + n_s) \cdot \sigma_{\epsilon_\eta}^2} = 1 + o_p(1) \qquad \textit{and} \qquad \frac{\widehat{\boldsymbol{S}}_{S\alpha}^T \widehat{\boldsymbol{S}}_{S\alpha}}{v_{S\alpha}} = 1 + o_p(1),$$

where

$$\boldsymbol{y}_\eta^T \boldsymbol{y}_\eta = \big(\boldsymbol{W}_{(1,\eta)} \boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta w}\big)^T \big(\boldsymbol{W}_{(1,\eta)} \boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta w}\big) + \big(\boldsymbol{S}_{(1,\eta)} \boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta s}\big)^T \big(\boldsymbol{S}_{(1,\eta)} \boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_{\eta s}\big),$$

$$
\begin{aligned}
\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha} =&\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)^T\boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)+\\
&2\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)^T\boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)+\\
&\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)^T\boldsymbol{S}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)+\\
&\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)^T\boldsymbol{X}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)+\\
&2\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)^T\boldsymbol{X}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)+\\
&\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big),
\end{aligned}
$$

and

$$
\begin{aligned}
v_{S\alpha} =&\{n_1n_3m_\alpha(p+n_1)\cdot\sigma_\alpha^2/p+n_1n_3p\cdot\sigma_{\epsilon_\alpha}^2\}+2\{n_1n_3n_sm_\alpha\cdot\sigma_\alpha^2/p\}+\\
&\{n_sn_3m_\alpha(p+n_s)\cdot\sigma_\alpha^2/p+n_sn_3p\cdot\sigma_{\epsilon_\alpha}^2\}+\{n_1n_sm_\alpha(p+n_1)\cdot\sigma_\alpha^2/p+n_1n_sp\cdot\sigma_{\epsilon_\alpha}^2\}\\
&+2\{n_1n_sm_\alpha(n_s+p)\cdot\sigma_\alpha^2/p\}+\{n_sm_\alpha(n_s^2+p^2+3n_sp)\cdot\sigma_\alpha^2/p+n_sp(n_s+p)\cdot\sigma_{\epsilon_\alpha}^2\}.
\end{aligned}
$$

Further if $p/\{(n_1+n_S)(n_3+n_s)\}\to 0$*, then we have*

$$
\begin{aligned}
&\frac{\boldsymbol{y}_\eta^T\widehat{\boldsymbol{S}}_{S\alpha}}{(n_1+n_s)n_3m_{\alpha\eta}\cdot\sigma_{\alpha\eta}/p+\{n_sm_{\alpha\eta}(p+n_s)\cdot\sigma_{\alpha\eta}/p+n_sp\cdot\sigma_{\epsilon_\alpha\epsilon_\eta}\}+n_sn_1m_{\alpha\eta}\cdot\sigma_{\alpha\eta}/p}\\
&=1+o_p(1),
\end{aligned}
$$

where $\boldsymbol{y}_\eta^T\widehat{\boldsymbol{S}}_{S\alpha}$ *is given by*

$$
\begin{aligned}
&\big(\boldsymbol{\epsilon}_{\eta w}^T+\boldsymbol{\eta}_{(1)}^T\boldsymbol{W}_{(1,\eta)}^T\big)\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big)+\big(\boldsymbol{\epsilon}_{\eta w}^T+\boldsymbol{\eta}_{(1)}^T\boldsymbol{W}_{(1,\eta)}^T\big)\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)\\
&+\big(\boldsymbol{\epsilon}_{\eta s}^T+\boldsymbol{\eta}_{(1)}^T\boldsymbol{S}_{(1,\eta)}^T\big)\boldsymbol{S}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha s}\big)+\big(\boldsymbol{\epsilon}_{\eta s}^T+\boldsymbol{\eta}_{(1)}^T\boldsymbol{S}_{(1,\eta)}^T\big)\boldsymbol{S}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_{\alpha x}\big).
\end{aligned}
$$

Proposition A.7. *Under polygenic model (3.4) and Conditions 3.1, 3.2, and 3.4, suppose*

$m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$ as $(n_1+n_s),(n_2+n_s),n_3,p \to \infty$, then we have

$$\frac{\widehat{\boldsymbol{S}}_{S\alpha}^T \widehat{\boldsymbol{S}}_{S\alpha}}{v_{S\alpha}} = 1 + o_p(1) \qquad \textit{and} \qquad \frac{\widehat{\boldsymbol{S}}_{S\beta}^T \widehat{\boldsymbol{S}}_{S\beta}}{v_{S\beta}} = 1 + o_p(1),$$

where

$$\begin{aligned}\widehat{\boldsymbol{S}}_{S\alpha}^T \widehat{\boldsymbol{S}}_{S\alpha} =& \big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha x}\big)^T \boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha x}\big) + \\ & 2\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha x}\big)^T \boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha s}\big) + \\ & \big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha s}\big)^T \boldsymbol{S}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha s}\big),\end{aligned}$$

$$\begin{aligned}\widehat{\boldsymbol{S}}_{S\beta}^T \widehat{\boldsymbol{S}}_{S\beta} =& \big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta z}\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta z}\big) + \\ & 2\big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta z}\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta s}\big) + \\ & \big(\boldsymbol{S}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta s}\big)^T \boldsymbol{S}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta s}\big),\end{aligned}$$

$$\begin{aligned}v_{S\alpha} =& n_1 n_3 m_\alpha (p+n_1)\cdot\sigma_\alpha^2/p + n_1 n_3 p\cdot\sigma_{\epsilon_\alpha}^2 + 2n_1 n_3 n_s m_\alpha\cdot\sigma_\alpha^2/p + \\ & n_s n_3 m_\alpha (p+n_s)\cdot\sigma_\alpha^2/p + n_s n_3 p\cdot\sigma_{\epsilon_\alpha}^2,\end{aligned}$$

and

$$\begin{aligned}v_{S\beta} =& n_2 n_3 m_\beta (p+n_2)\cdot\sigma_\beta^2/p + n_2 n_3 p\cdot\sigma_{\epsilon_\beta}^2 + 2n_2 n_3 n_s m_\beta\cdot\sigma_\beta^2/p + \\ & n_s n_3 m_\beta (p+n_s)\cdot\sigma_\beta^2/p + n_s n_3 p\cdot\sigma_{\epsilon_\beta}^2.\end{aligned}$$

Further if $p^2/\{(n_1+n_s)(n_2+n_s)n_3\} \to 0$, then we have

$$\frac{\widehat{\boldsymbol{S}}_{S\alpha}^T \widehat{\boldsymbol{S}}_{S\beta}}{v_{S\alpha\beta}} = 1 + o_p(1),$$

where

$$\begin{aligned}\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\beta} = &\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha x}\big)^T\boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta z}\big)+\\ &\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha x}\big)^T\boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta s}\big)+\\ &\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha s}\big)^T\boldsymbol{S}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta z}\big)+\\ &\big(\boldsymbol{S}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_{\alpha s}\big)^T\boldsymbol{S}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{S}^T\big(\boldsymbol{S}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_{\beta s}\big),\end{aligned}$$

and

$$\begin{aligned}v_{S\alpha\beta} =&n_1n_2n_3m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p + n_1n_sn_3m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p + n_sn_2n_3m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p+\\ &\{n_s(p+n_s)n_3m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p + n_sn_3p\cdot\sigma_{\epsilon_\alpha\epsilon_\beta}\}.\end{aligned}$$

Then Theorem 3.4 and Proposition A.4 follow from continuous mapping theorem and similar arguments above.

More overlapping cases

This section provides more analyses on the overlapping samples. We consider several additional cases that might occur in real data applications.

Case ii)

In this case, we add n_s overlapping samples into Discovery GWAS-I and II, resulting in the following two new datasets:

- Dataset IV: $(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{y}_\alpha)$, with $\boldsymbol{X} \in \mathbb{R}^{n_1\times p}$, $\boldsymbol{S} \in \mathbb{R}^{n_s\times p}$, and $\boldsymbol{y}_\alpha^T = (\boldsymbol{y}_{\alpha_X}^T, \boldsymbol{y}_{\alpha_S}^T) \in \mathbb{R}^{(n_1+n_s)\times 1}$.
- Dataset VI: $(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{y}_\beta)$, with $\boldsymbol{Z} \in \mathbb{R}^{n_2\times p}$, $\boldsymbol{S} \in R^{n_s\times p}$, and $\boldsymbol{y}_\beta^T = (\boldsymbol{y}_{\beta_Z}^T, \boldsymbol{y}_{\beta_S}^T) \in \mathbb{R}^{(n_2+n_s)\times 1}$.

Then we define $h_{\alpha\beta} \in (0, 1]$ as

$$h_{\alpha\beta} = \frac{(m_{\alpha\beta}/p)\sigma_{\alpha\beta}}{(m_{\alpha\beta}/p)\sigma_{\alpha\beta} + \sigma_{\epsilon_\alpha\epsilon_\beta}},$$

which quantifies the contribution of genetic correlation to the phenotypic correlation. We introduce the following additional condition on random errors.

Condition A.1. *On n_s overlapping samples, ϵ_{α_j} and ϵ_{β_j} are independent random variables satisfying*

$$\begin{pmatrix} \epsilon_{\alpha_j} \\ \epsilon_{\beta_j} \end{pmatrix} \sim F \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\epsilon_\alpha} & \sigma_{\epsilon_\alpha \epsilon_\beta} \\ \sigma_{\epsilon_\alpha \epsilon_\beta} & \sigma^2_{\epsilon_\beta} \end{pmatrix} \right]$$

for $j = 1, ..., n_s$, where $\sigma_{\epsilon_\alpha \epsilon_\beta} = \rho_{\epsilon_\alpha \epsilon_\beta} \cdot \sigma_{\epsilon_\alpha} \sigma_{\epsilon_\beta}$.

Proposition A.8. *Under polygenic models and Conditions 1, 2, and Condition S1, suppose $min(m_{\alpha\beta},\ m_\alpha,\ m_\beta) \to \infty$ as $min\{(n_1 + n_s), (n_2 + n_s), n_3, p\} \to \infty$, and let $p = c \cdot \{(n_1 + n_s)(n_2 + n_s)n_3\}^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then $G_{S\alpha\beta}$ is given by*

$$\frac{(n_1 + n_s)^{1/2}(n_2 + n_s)^{1/2} + n_s p / \{(n_1 + n_s)^{1/2}(n_2 + n_s)^{1/2} \cdot h_{\alpha\beta}\}}{\left\{(n_1 + n_s + p/h^2_\alpha) \cdot (n_2 + n_s + p/h^2_\beta)\right\}^{1/2}} \cdot \varphi_{\alpha\beta} + o_p(1).$$

If $a \in [1, \infty]$, then we have $G_{S\alpha\beta} = o_p(1)$.

Proposition A.8 shows the effect of n_s overlapping samples on the estimation of $\varphi_{\alpha\beta}$. When the two discovery GWAS are fully overlapped, i.e., the two set of summary statistics are generated from the same GWAS, then we have

$$G_{S\alpha\beta} = \frac{n_s + p/h_{\alpha\beta}}{\left\{(n_s + p/h^2_\alpha) \cdot (n_s + p/h^2_\beta)\right\}^{1/2}} \cdot \varphi_{\alpha\beta} + o_p(1).$$

In the optimal situation with $h^2_\alpha = h^2_\beta = h_{\alpha\beta} = 1$, we have $G_{S\alpha\beta} = \varphi_{\alpha\beta} + o_p(1)$. Thus, $G_{S\alpha\beta}$ is a consistent estimator and we may have an unbiased estimator of genetic correlation.

Case iii)

When the two GWAS are fully overlapped, i.e., the two set of summary statistics $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are generated from the same GWAS data

- Dataset VII: $(\boldsymbol{X}, \boldsymbol{y}_\alpha, \boldsymbol{y}_\beta)$, with $\boldsymbol{X} = [\boldsymbol{X}_{(1,\alpha)}, \boldsymbol{X}_{(2,\alpha)}] = [\boldsymbol{X}_{(1,\beta)}, \boldsymbol{X}_{(2,\beta)}] \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{X}_{(1,\alpha)} \in \mathbb{R}^{n_1 \times m_\alpha}$, $\boldsymbol{X}_{(1,\beta)} \in \mathbb{R}^{n_1 \times m_\beta}$, $\boldsymbol{y}_\alpha \in \mathbb{R}^{n_1 \times 1}$, and $\boldsymbol{y}_\beta \in \mathbb{R}^{n_1 \times 1}$.

We assume that $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ have polygenic architectures. We estimate $\varphi_{\alpha\beta}$ directly by estimating the correlation of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$

$$\widehat{\varphi}_{X\alpha\beta} = \frac{\widehat{\boldsymbol{\alpha}}^T \widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\alpha}}\| \cdot \|\widehat{\boldsymbol{\beta}}\|} = \frac{\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{\big\{\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\big(\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{X}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1,\beta)}\beta_{(1)} + \boldsymbol{\epsilon}_\beta\big)\big\}\big]^{1/2}}.$$

Proposition A.9. *Under polygenic model and Conditions 3.1, 3.2, and A.1, if $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$, as $n_1, p \to \infty$, then we have*

$$\frac{\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)}{\{n_1 m_\alpha(n_1 + m_\alpha) + n_1 m_\alpha(p - m_\alpha)\} \cdot \sigma_\alpha^2/p + n_1 p \cdot \sigma_{\epsilon_\alpha}^2} = 1 + o_p(1),$$
$$\frac{\big(\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{X}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{\{n_1 m_\beta(n_1 + m_\beta) + n_1 m_\beta(p - m_\beta)\} \cdot \sigma_\beta^2/p + n_1 p \cdot \sigma_{\epsilon_\beta}^2} = 1 + o_p(1),$$

and

$$\frac{\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)}{n_1 m_{\alpha\beta}(p + n_1) \cdot \sigma_{\alpha\beta}/p + n_1 p \cdot \sigma_{\epsilon_{\alpha\beta}}} = 1 + o_p(1).$$

Thus, we have

$$\widehat{\varphi}_{X\alpha\beta} = \frac{n_1 + p/h_{\alpha\beta}}{(n_1 + p/h_\alpha^2)^{1/2}(n_1 + p/h_\beta^2)^{1/2}} \cdot \varphi_{\alpha\beta} + o_p(1).$$

It follows that $\widehat{\varphi}_{X\alpha\beta}$ is asymptotically unbiased as $h_\alpha^2 = h_\beta^2 = h_{\alpha\beta} = 1$. Otherwise, $\widehat{\varphi}_{X\alpha\beta}$ may be biased towards zero.

Case iv)

Again, the two set of summary statistics $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are generated from the same GWAS dataset

- Dataset VII: $(\boldsymbol{X}, \boldsymbol{y}_\alpha, \boldsymbol{y}_\beta)$, with $\boldsymbol{X} = [\boldsymbol{X}_{(1,\alpha)}, \boldsymbol{X}_{(2,\alpha)}] = [\boldsymbol{X}_{(1,\beta)}, \boldsymbol{X}_{(2,\beta)}] \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{X}_{(1,\alpha)} \in \mathbb{R}^{n_1 \times m_\alpha}$, $\boldsymbol{X}_{(1,\beta)} \in \mathbb{R}^{n_1 \times m_\beta}$, $\boldsymbol{y}_\alpha \in \mathbb{R}^{n_1 \times 1}$, and $\boldsymbol{y}_\beta \in \mathbb{R}^{n_1 \times 1}$.

And we construct two PRSs $\widehat{\boldsymbol{S}}_{X\alpha}$ and $\widehat{\boldsymbol{S}}_{X\beta}$ on $\boldsymbol{X}$.

Proposition A.10. *Under polygenic model and Conditions 3.1, 3.2, and A.1, if $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$, as $n_1, p \to \infty$, then we have*

$$\frac{\widehat{\boldsymbol{S}}_{X\alpha}^T \widehat{\boldsymbol{S}}_{X\alpha}}{v_{X\alpha}} = 1 + o_p(1) \qquad \textit{and} \qquad \frac{\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\beta}}{v_{X\beta}} = 1 + o_p(1),$$

where

$$\begin{aligned}
&\widehat{\boldsymbol{S}}_{X\alpha}^T \widehat{\boldsymbol{S}}_{X\alpha} = (\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha)^T \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha),\\
&\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\beta} = (\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta)^T \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta),\\
&v_{X\alpha} = n_1 m_\alpha\{(n_1+p)^2 + n_1 p\} \cdot \sigma_\alpha^2/p + n_1 p(n_1+p) \cdot \sigma_{\epsilon_\alpha}^2, \quad \textit{and}\\
&v_{X\beta} = n_1 m_\beta\{(n_1+p)^2 + n_1 p\} \cdot \sigma_\beta^2/p + n_1 p(n_1+p) \cdot \sigma_{\epsilon_\beta}^2.
\end{aligned}$$

Similarly, we have

$$\frac{\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\alpha}}{v_{X\alpha\beta}} = 1 + o_p(1),$$

where $v_{X\alpha\beta} = n_1 m_{\alpha\beta}\{(n_1+p)^2 + n_1 p\} \cdot \sigma_{\alpha\beta}/p + n_1 p(n_1+p) \cdot \sigma_{\epsilon_{\alpha\beta}}$ *and*

$$\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\alpha} = (\boldsymbol{X}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta)^T \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha).$$

Thus, we have

$$G_{X\alpha\beta} = \frac{n_1^2 + 2n_1 p + p(n_1+p)/h_{\alpha\beta}}{\left\{n_1^2 + 2n_1 p + p(n_1+p)/h_\alpha^2\right\}^{1/2}\left\{n_1^2 + 2n_1 p + p(n_1+p)/h_\beta^2\right\}^{1/2}} \cdot \varphi_{\alpha\beta} + o_p(1).$$

It follows that $G_{X\alpha\beta}$ is asymptotically unbiased if $h_\alpha^2 = h_\beta^2 = h_{\alpha\beta} = 1$. Otherwise, $G_{X\alpha\beta}$ may

be biased towards zero.

Case v)

The two set of GWAS summary statistics $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are generated from the following two independent datasets:

- Dataset VIII: $(\boldsymbol{X}, \boldsymbol{y}_\alpha)$, with $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{X}_{(1)} \in \mathbb{R}^{n_1 \times m_\alpha}$, and $\boldsymbol{y}_\alpha \in \mathbb{R}^{n_1 \times 1}$.
- Dataset IX: $(\boldsymbol{Z}, \boldsymbol{y}_\beta)$, with $\boldsymbol{Z} = [\boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)}] \in \mathbb{R}^{n_2 \times p}$, $\boldsymbol{Z}_{(1)} \in \mathbb{R}^{n_2 \times m_\beta}$, and $\boldsymbol{y}_\beta \in \mathbb{R}^{n_2 \times 1}$.

We construct two PRSs $\widehat{\boldsymbol{S}}_{X\alpha}$ and $\widehat{\boldsymbol{S}}_{X\beta}$ on $\boldsymbol{X}$.

Proposition A.11. *Under polygenic model and Conditions 3.1, 3.2, and A.1, if $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$, as $n_1, n_2, p \to \infty$, then we have*

$$\frac{\widehat{\boldsymbol{S}}_{X\alpha}^T \widehat{\boldsymbol{S}}_{X\alpha}}{v_{X\alpha}} = 1 + o_p(1) \qquad \textit{and} \qquad \frac{\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\beta}}{v_{X\beta}} = 1 + o_p(1),$$

where

$$\begin{aligned}
&\widehat{\boldsymbol{S}}_{X\alpha}^T \widehat{\boldsymbol{S}}_{X\alpha} = \big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big),\\
&\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\beta} = \big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big),\\
&v_{X\alpha} = n_1 m_\alpha\{(n_1+p)^2 + n_1 p\} \cdot \sigma_\alpha^2/p + n_1 p(n_1+p) \cdot \sigma_{\epsilon_\alpha}^2, \quad \textit{and}\\
&v_{X\beta} = n_1 n_2 m_\beta (p+n_2) \cdot \sigma_\beta^2/p + n_1 n_2 p \cdot \sigma_{\epsilon_\beta}^2.
\end{aligned}$$

Further if $p/(n_1 n_2) \to 0$, then we have

$$\frac{\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\alpha}}{v_{X\alpha\beta}} = 1 + o_p(1),$$

where $v_{X\alpha\beta} = n_1 n_2 m_{\alpha\beta}(n_1+p) \cdot \sigma_{\alpha\beta}/p$ *and*

$$\widehat{\boldsymbol{S}}_{X\beta}^T \widehat{\boldsymbol{S}}_{X\alpha} = \big(\boldsymbol{X}_{(1,\alpha)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1,\beta)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big).$$

Let $p = c \cdot (n_1 n_2)^a$ for some constants $c > 0$ and $a \in (0, \infty]$. As n_1 and n_2 increase to ∞, if $a \in (0, 1)$, then we have

$$G_{X\alpha\beta} = \frac{(n_1 + p) \cdot n_2^{1/2}}{\left\{n_1^2 + 2n_1 p + p(n_1 + p)/h_\alpha^2\right\}^{1/2} \left\{n_2 + p/h_\beta^2\right\}^{1/2}} \cdot \varphi_{\alpha\beta} + o_p(1).$$

Variance of cross-trait PRS

In this section, we study the variance of cross-trait PRS. Consider

$$G_{\alpha\eta} = \frac{\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_\alpha}{\left\|\boldsymbol{y}_\eta\right\| \cdot \left\|\widehat{\boldsymbol{S}}_\alpha\right\|} \quad \text{and} \quad G_{\alpha\eta}^2 = \frac{(\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_\alpha)^2}{\left\|\boldsymbol{y}_\eta\right\|^2 \cdot \left\|\widehat{\boldsymbol{S}}_\alpha\right\|^2},$$

where

$$\left\|\boldsymbol{y}_\eta\right\|^2 = \boldsymbol{y}_\eta^T \boldsymbol{y}_\eta = \left(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\right)^T \left(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\right) = \left(n_3 m_\eta \cdot \sigma_\eta^2/p + n_3 \cdot \sigma_{\epsilon_\eta}^2\right) \cdot \{1 + o_p(1)\},$$

and

$$\begin{aligned}
\left\|\widehat{\boldsymbol{S}}_\alpha\right\|^2 = \widehat{\boldsymbol{S}}_\alpha^T \widehat{\boldsymbol{S}}_\alpha &= \left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\right)^T \boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T \left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\right) \\
&= \left[\left\{n_1 n_3 m_\alpha (n_1 + p)\right\} \cdot \sigma_\alpha^2/p + n_1 n_3 p \cdot \sigma_{\epsilon_\alpha}^2\right] \cdot \{1 + o_p(1)\}
\end{aligned}$$

When $p = O(n_1)$, we have $p/(n_1 n_3) \to 0$ as $\min(n_1, n_3, p) \to \infty$, and thus

$$\begin{aligned}
(\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_\alpha)^2 &= \left\{\left(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\right)^T \boldsymbol{W}\boldsymbol{X}^T \left(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\right)\right\}^2 \\
&= \left\{(n_1^2 n_3^2 m_{\alpha\eta}^2 + n_1 n_3 m_{\alpha\eta}^2 p + 2n_1^2 n_3 m_{\alpha\eta}^2 + 2n_1 n_3^2 m_{\alpha\eta}^2) \cdot \sigma_{\alpha\eta}^2 + \right. \\
&\qquad \left. n_1^2 n_3^2 m_{\alpha\eta} \cdot (\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2)\right\} p^{-2} \cdot \{1 + o_p(1)\}.
\end{aligned}$$

It follows that

$$G_{\alpha\eta}^2 = \frac{(n_1n_3 + p + 2n_1 + 2n_3) \cdot m_{\alpha\eta}^2\sigma_{\alpha\eta}^2 + n_1n_3m_{\alpha\eta} \cdot (a_{22} - \sigma_{\alpha\eta}^2)}{\big(n_3 \cdot m_\eta\sigma_\eta^2/h_\eta^2\big) \cdot \{(p/h_\alpha^2 + n_1) \cdot m_\alpha\sigma_\eta^2\}} + o_p(1)$$
$$= \Big\{\frac{n_1n_3 + p + 2n_1 + 2n_3}{(n_3/h_\eta^2) \cdot (p/h_\alpha^2 + n_1)} \cdot \varphi_{\alpha\eta}^2 + \frac{n_1n_3}{(n_3/h_\eta^2) \cdot (p/h_\alpha^2 + n_1)} \cdot \frac{m_{\alpha\eta}(a_{22} - \sigma_{\alpha\eta}^2)}{m_\alpha m_\eta\sigma_\alpha^2\sigma_\eta^2}\Big\} + o_p(1).$$

Since $G_{\alpha\eta} \in [0, 1]$, we have $\mathrm{E}(G_{\alpha\eta}) = \{n_1(n_1 + p/h_\alpha^2)\}^{1/2} \cdot h_\eta \cdot \varphi_{\alpha\eta}$ and

$$\mathrm{Var}(G_{\alpha\eta}) = \mathrm{E}(G_{\alpha\eta}^2) - \{\mathrm{E}(G_{\alpha\eta})\}^2$$
$$= \Big\{\frac{p + 2n_1 + 2n_3}{n_3(p/h_\alpha^2 + n_1)} \cdot h_\eta^2 \cdot \varphi_{\alpha\eta}^2 + \frac{n_1}{p/h_\alpha^2 + n_1} \cdot h_\eta^2 \cdot \frac{m_{\alpha\eta}(a_{22} - \sigma_{\alpha\eta}^2)}{m_\alpha m_\eta\sigma_\alpha^2\sigma_\eta^2}\Big\} \cdot \{1 + o_p(1)\}.$$

In addition, we have

$$T^2 = \frac{\{\mathrm{E}(G_{\alpha\eta})\}^2}{\mathrm{Var}(G_{\alpha\eta})} = \Big\{\frac{p + 2n_1 + 2n_3}{n_1n_3} + \frac{m_{\alpha\eta}(\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2)}{m_\alpha m_\eta\sigma_\alpha^2\sigma_\eta^2\varphi_{\alpha\eta}^2}\Big\}^{-1} \cdot \{1 + o_p(1)\}$$
$$= \Big\{\frac{p + 2n_1 + 2n_3}{n_1n_3} + \frac{\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2}{m_{\alpha\eta}\sigma_{\alpha\eta}^2}\Big\}^{-1} \cdot \{1 + o_p(1)\},$$

where $\mathrm{E}(\alpha_1^2\eta_1^2) = \sigma_{\alpha^2\eta^2}/p^2$, and $\mathrm{E}(\alpha_1\eta_1) = \sigma_{\alpha\eta}/p$. Then Corollary 3.1 holds. Similarly, consider

$$G_{\alpha\beta} = \frac{\widehat{\boldsymbol{S}}_\beta^T\widehat{\boldsymbol{S}}_\alpha}{\big\|\widehat{\boldsymbol{S}}_\beta\big\| \cdot \big\|\widehat{\boldsymbol{S}}_\alpha\big\|} \quad \text{and} \quad G_{\alpha\beta}^2 = \frac{(\widehat{\boldsymbol{S}}_\beta^T\widehat{\boldsymbol{S}}_\alpha)^2}{\big\|\widehat{\boldsymbol{S}}_\beta\big\|^2 \cdot \big\|\widehat{\boldsymbol{S}}_\alpha\big\|^2},$$

where

$$\big\|\widehat{\boldsymbol{S}}_\alpha\big\|^2 = \widehat{\boldsymbol{S}}_\alpha^T\widehat{\boldsymbol{S}}_\alpha = \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T\boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)$$
$$= \Big[\big\{n_1n_3m_\alpha(n_1 + p)\big\} \cdot \sigma_\alpha^2/p + n_1n_3p \cdot \sigma_{\epsilon_\alpha}^2\Big] \cdot \{1 + o_p(1)\},$$

and

$$\begin{aligned}\big\|\widehat{\boldsymbol{S}}_\beta\big\|^2 = \widehat{\boldsymbol{S}}_\beta^T\widehat{\boldsymbol{S}}_\beta &= \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}_\beta\big)^T\boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}_\beta\big)\\ &= \Big[\big\{n_2n_3m_\beta(n_2+p)\big\}\cdot\sigma_\beta^2/p + n_2n_3p\cdot\sigma_{\epsilon_\beta}^2\Big]\cdot\{1+o_p(1)\}.\end{aligned}$$

When $p = O(n_1) = O(n_2)$, as $\min(n_1, n_2, n_3, p) \to \infty$, we have

$$\begin{aligned}(\widehat{\boldsymbol{S}}_\beta^T\widehat{\boldsymbol{S}}_\alpha)^2 &= \big\{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}_\beta\big)^T\boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_\alpha\big)\big\}^2\\ &= \big\{O(n_1n_2n_3m_{\alpha\beta}^2p^2 + n_1^2n_2^2n_3^2m_{\alpha\beta}) + n_1^2n_2^2n_3^2m_{\alpha\beta}^2\cdot\sigma_{\alpha\beta}^2\big\}p^{-2}\cdot\{1+o_p(1)\}.\end{aligned}$$

It follows that

$$\begin{aligned}G_{\alpha\beta}^2 &= \frac{O(n_3^{-1}m_{\alpha\beta}^2p^2 + n_1n_2m_{\alpha\beta}) + n_1n_2m_{\alpha\beta}^2\cdot\sigma_{\alpha\beta}^2}{\{(p/h_\beta^2+n_2)\cdot m_\beta\sigma_\beta^2\}\cdot\{(p/h_\alpha^2+n_1)\cdot m_\alpha\sigma_\alpha^2\}}\cdot\{1+o_p(1)\}\\ &= \Big[O\Big\{\frac{n_3^{-1}p^2+n_1n_2m_{\alpha\beta}^{-1}}{(p/h_\beta^2+n_2)\cdot(p/h_\alpha^2+n_1)}\Big\} + \frac{n_1n_2}{(p/h_\beta^2+n_2)\cdot(p/h_\alpha^2+n_1)}\cdot\varphi_{\alpha\beta}^2\Big]\cdot\{1+o_p(1)\},\end{aligned}$$

and

$$\mathrm{Var}(G_{\alpha\beta}) = \mathrm{E}(G_{\alpha\beta}^2) - \{\mathrm{E}(G_{\alpha\beta})\}^2 = O\Big\{\frac{n_3^{-1}p^2+n_1n_2m_{\alpha\beta}^{-1}}{(p/h_\beta^2+n_2)\cdot(p/h_\alpha^2+n_1)}\Big\}\cdot\{1+o_p(1)\}.$$

Now consider

$$\widehat{\varphi}_{\alpha\beta} = \frac{\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\alpha}}\|\cdot\|\widehat{\boldsymbol{\beta}}\|} \quad\text{and}\quad \widehat{\varphi}_{\alpha\beta}^2 = \frac{(\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\beta}})^2}{\|\widehat{\boldsymbol{\alpha}}\|^2\cdot\|\widehat{\boldsymbol{\beta}}\|^2},$$

where

$$\begin{aligned}\big\|\widehat{\boldsymbol{\beta}}\big\|^2 &= \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}_\beta\big)^T\boldsymbol{Z}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}_\beta\big)\\ &= \Big[\big\{n_2m_\beta(n_2+p)\big\}\cdot\sigma_\beta^2/p + n_2p\cdot\sigma_{\epsilon_\beta}^2\Big]\cdot\{1+o_p(1)\}\end{aligned}$$

and

$$\begin{aligned}\big\|\widehat{\boldsymbol{\alpha}}\big\|^2 &= \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \\ &= \Big[\big\{n_1 m_\alpha (n_1+p)\big\}\cdot \sigma_\alpha^2/p + n_1 p\cdot \sigma_{\epsilon_\alpha}^2\Big]\cdot\{1+o_p(1)\}.\end{aligned}$$

When $p = O(n_1) = O(n_2)$, as $\min(n_1, n_2, p) \to \infty$, we have

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\beta}})^2 &= \big\{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\big\}^2 \\ &= \big\{O(n_1 n_2 m_{\alpha\beta}^2 p + n_1^2 n_2^2 m_{\alpha\beta}) + n_1^2 n_2^2 m_{\alpha\beta}^2\cdot\sigma_{\alpha\beta}^2\big\}p^{-2}\cdot\{1+o_p(1)\}.\end{aligned}$$

It follows that

$$\begin{aligned}\widehat{\varphi}_{\alpha\beta}^2 &= \frac{O(m_{\alpha\beta}^2 p + n_1 n_2 m_{\alpha\beta}) + n_1 n_2 m_{\alpha\beta}^2\cdot\sigma_{\alpha\beta}^2}{\{(p/h_\beta^2+n_2)\cdot m_\beta\sigma_\beta^2\}\cdot\{(p/h_\alpha^2+n_1)\cdot m_\alpha\sigma_\alpha^2\}}\cdot\{1+o_p(1)\} \\ &= \Big[O\Big\{\frac{p+n_1n_2m_{\alpha\beta}^{-1}}{(p/h_\beta^2+n_2)\cdot(p/h_\alpha^2+n_1)}\Big\} + \frac{n_1n_2}{(p/h_\beta^2+n_2)\cdot(p/h_\alpha^2+n_1)}\cdot\varphi_{\alpha\beta}^2\Big]\cdot\{1+o_p(1)\},\end{aligned}$$

and

$$\mathrm{Var}(\widehat{\varphi}_{\alpha\beta}) = \mathrm{E}(\widehat{\varphi}_{\alpha\beta}^2) - \{\mathrm{E}(\widehat{\varphi}_{\alpha\beta})\}^2 = O\Big\{\frac{p+n_1n_2m_{\alpha\beta}^{-1}}{(p/h_\beta^2+n_2)\cdot(p/h_\alpha^2+n_1)}\Big\}\cdot\{1+o_p(1)\}.$$

Intermediate results: cross-trait PRS with all SNPs

Proposition A.12. *Under polygenic model and Conditions 3.1, 3.2, if* $m_{\alpha\eta}, m_\alpha$, *and* $m_\eta \to \infty$ *as* $n_1, n_3, p \to \infty$, *then we have*

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \boldsymbol{W}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \Big\} = n_1 n_3 m_{\alpha\eta} \cdot \sigma_{\alpha\eta}/p, \\
&Var\Big\{ \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \boldsymbol{W}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \Big\} = \{(n_1 n_3 m_{\alpha\eta}^2 p + \\
&\quad 2n_1^2 n_3 m_{\alpha\eta}^2 + 2n_1 n_3^2 m_{\alpha\eta}^2) \cdot \sigma_{\alpha\eta}^2 + n_1^2 n_3^2 m_{\alpha\eta} \cdot (\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2)\} p^{-2} \cdot \{1 + o(1)\},
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big) \Big\} = n_3 m_\eta \cdot \sigma_\eta^2/p + n_3 \cdot \sigma_{\epsilon_\eta}^2, \\
&Var\Big\{ \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big)^T \big(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\big) \Big\} = o(n_3^2 m_\eta^2 \cdot \sigma_\eta^4/p^2),
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \Big\} \\
&= \big[n_1 n_3 m_\alpha (n_1 + m_\alpha) \cdot \{1 + o(1)\} + n_1 n_3 m_\alpha (p - m_\alpha)\big] \cdot \sigma_\alpha^2/p + n_1 n_3 p \cdot \sigma_{\epsilon_\alpha}^2, \\
&Var\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \Big\} = o\{n_1^2 n_3^2 m_\alpha^2 (n_1 + p)^2 \cdot \sigma_\alpha^4/p^2\},
\end{aligned}
$$

where $E(\alpha^2\eta^2) = \sigma_{\alpha^2\eta^2}/p^2$.

Proposition A.13. *Under polygenic model and Conditions 3.1, 3.2, if* $m_{\alpha\beta}, m_\alpha$, *and* $m_\beta \to \infty$ *as* $n_1, n_2, n_3, p \to \infty$, *then we have*

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \Big\} = n_1 n_2 n_3 m_{\alpha\beta} \cdot \sigma_{\alpha\beta}/p, \\
&Var\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{X}^T \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big) \Big\} \\
&= O(n_1 n_2 n_3 m_{\alpha\beta}^2) + o(n_1^2 n_2^2 n_3^2 m_{\alpha\beta}^2/p^2),
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)\Big\} \\
&= \big[n_2 n_3 m_\beta (n_2 + m_\beta)\cdot\{1+o(1)\} + n_2 n_3 m_\beta (p - m_\beta)\big]\cdot \sigma_\beta^2/p + n_2 n_3 p\cdot\sigma_{\epsilon_\beta}^2, \\
&Var\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)\Big\} = o\{n_2^2 n_3^2 m_\beta^2 (n_2+p)^2\cdot\sigma_\beta^4/p^2\}.
\end{aligned}
$$

Proposition A.14. *Under polygenic model and Conditions 3.1, 3.2, if $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$ as $n_1, n_2, p \to \infty$, then we have*

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\Big\} = n_1 n_2 m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p, \\
&Var\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\Big\} = O(n_1 n_2 m_{\alpha\beta}^2/p) + o(n_1^2 n_2^2 m_{\alpha\beta}^2/p^2),
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\Big\} \\
&= \big[n_1 m_\alpha (n_1 + m_\alpha)\cdot\{1+o(1)\} + n_1 m_\alpha (p - m_\alpha)\big]\cdot \sigma_\alpha^2/p + n_1 p\cdot\sigma_{\epsilon_\alpha}^2, \\
&Var\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)^T \boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha\big)\Big\} = o\{n_1^2 m_\alpha^2 (n_1+p)^2\cdot\sigma_\alpha^4/p^2\},
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)\Big\} \\
&= \big[n_2 m_\beta (n_2 + m_\beta)\cdot\{1+o(1)\} + n_2 m_\beta (p - m_\beta)\big]\cdot \sigma_\beta^2/p + n_2 p\cdot\sigma_{\epsilon_\beta}^2, \\
&Var\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)^T \boldsymbol{Z}\boldsymbol{Z}^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta\big)\Big\} = o\{n_2^2 m_\beta^2 (n_2+p)^2\cdot\sigma_\beta^4/p^2\}.
\end{aligned}
$$

Then Propositions A.1 - A.3 follow from Markov's inequality.

Intermediate results: cross-trait PRS with selected SNPs

Proposition A.15. *Under polygenic model and Conditions 3.1, 3.2, suppose $m_{\alpha\eta}$, $m_{\alpha\beta}$, m_α, m_η, and $m_\beta \to \infty$, $q_{\alpha\beta}, q_{\alpha1}, q_{\alpha2}, q_{\beta1}, q_{\beta2}$, and $q_{\alpha\eta} \to \infty$ as $n_1, n_2, n_3, p \to \infty$, then we*

have

$$E(C_{\alpha\eta}) = n_1 n_3 q_{\alpha\eta} \cdot \sigma_{\alpha\eta}/p,$$
$$Var(C_{\alpha\eta}) = O(m_{\alpha\eta}^2 n_1 n_3 q_\alpha/p^2) + o(n_1^2 n_3^2 q_{\alpha\eta}^2/p^2),$$

$$E(V_\alpha) = \{n_1 n_3 m_\alpha q_{\alpha 2} + n_1 n_3 q_{\alpha 1}(m_\alpha + n_1)\} \cdot \sigma_\alpha^2/p \cdot \{1 + o(1)\} + n_1 n_3 q_\alpha \cdot \sigma_{\epsilon_\alpha}^2,$$
$$Var(V_\alpha) = o\big[\{n_1 n_3 m_\alpha q_{\alpha 2} + n_1 n_3 q_{\alpha 1}(m_\alpha + n_1)\}^2/p^2\big],$$

$$E(C_{\alpha\beta}) = n_1 n_2 n_3 q_{\alpha\beta} \cdot \sigma_{\alpha\beta}/p,$$
$$Var(C_{\alpha\eta}) = O(m_{\alpha\beta}^2 n_1 n_2 n_3 q_\alpha q_\beta/p^2) + o(n_1^2 n_2^2 n_3^2 q_{\alpha\beta}^2/p^2),$$

$$E(V_\beta) = \{n_2 n_3 m_\beta q_{\beta 2} + n_2 n_3 q_{\beta 1}(m_\beta + n_2)\} \cdot \sigma_\beta^2/p \cdot \{1 + o(1)\} + n_2 n_3 q_\alpha \cdot \sigma_{\epsilon_\alpha}^2,$$
$$Var(V_\alpha) = o\big[\{n_2 n_3 m_\beta q_{\beta 2} + n_2 n_3 q_{\beta 1}(m_\beta + n_2)\}^2/p^2\big].$$

Then Propositions A.4 and A.5 follow from Markov's inequality.

Intermediate results: overlapping samples

Proposition A.16. *Under polygenic model and Conditions 3.1 - 3.3, suppose* $m_{\alpha\eta}, m_\alpha$, *and* $m_\eta \to \infty$ *as* $(n_1 + n_s), (n_3 + n_s), p \to \infty$, *then we have*

$$E\big(\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_{S\alpha}\big) = (n_3 + n_s)(n_1 + n_s) m_{\alpha\eta} \cdot \sigma_{\alpha\eta}/p + n_s m_{\alpha\eta} p \cdot \sigma_{\alpha\eta}/p + n_s p \cdot \sigma_{\epsilon_\alpha \epsilon_\eta},$$
$$Var\big(\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_{S\alpha}\big) = O\big\{(n_3 + n_s)(n_1 + n_s) m_{\alpha\eta}^2/p\big\} + o\big\{E^2(\boldsymbol{y}_\eta^T \widehat{\boldsymbol{S}}_{S\alpha})\big\},$$

$$E\big(\boldsymbol{y}_\eta^T\boldsymbol{y}_\eta\big) = (n_3+n_s)m_\eta\cdot\sigma_\eta^2/p + (n_3+n_s)\cdot\sigma_{\epsilon_\eta}^2,$$
$$Var\big(\boldsymbol{y}_\eta^T\boldsymbol{y}_\eta\big) = o\big\{(n_3+n_s)^2m_\eta^2\cdot\sigma_\eta^4/p^2\big\},$$

$$E\big(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha}\big) = \big\{n_1n_3m_\alpha(p+n_1)\cdot\sigma_\alpha^2/p + n_1n_3p\cdot\sigma_{\epsilon_\alpha}^2\big\} + 2\big\{n_1n_3n_sm_\alpha\cdot\sigma_\alpha^2/p\big\}+$$
$$\big\{n_sn_3m_\alpha(p+n_s)\cdot\sigma_\alpha^2/p + n_sn_3p\cdot\sigma_{\epsilon_\alpha}^2\big\} + \big\{n_1n_sm_\alpha(p+n_1)\cdot\sigma_\alpha^2/p + n_1n_sp\cdot\sigma_{\epsilon_\alpha}^2\big\}$$
$$+ 2\big\{n_1n_sm_\alpha(n_s+p)\cdot\sigma_\alpha^2/p\big\} + \big\{n_sm_\alpha(n_s^2+p^2+3n_sp)\cdot\sigma_\alpha^2/p + n_sp(n_s+p)\cdot\sigma_{\epsilon_\alpha}^2\big\},$$
$$Var\big(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha}\big) = o\big\{E^2(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha})\big\}.$$

Proposition A.17. *Under polygenic model and Conditions 3.1, 3.2, and A.1, suppose $m_{\alpha\beta}, m_\alpha$, and $m_\beta \to \infty$ as $(n_1+n_s),(n_2+n_s),n_3,p\to\infty$, then we have*

$$E\big(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\beta}\big) = (n_1+n_s)(n_2+n_s)n_3m_{\alpha\beta}\cdot\sigma_{\alpha\beta}/p + n_sn_3m_{\alpha\beta}p\cdot\sigma_{\alpha\beta}/p + n_sn_3p\cdot\sigma_{\epsilon_\alpha\epsilon_\beta},$$
$$Var\big(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\beta}\big) = O\big\{(n_1+n_s)(n_2+n_s)n_3m_{\alpha\eta}^2\big\} + o\big\{E^2(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\beta})\big\},$$

$$E(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha}) = n_1n_3m_\alpha(p+n_1)\cdot\sigma_\alpha^2/p + n_1n_3p\cdot\sigma_{\epsilon_\alpha}^2 + 2n_1n_3n_sm_\alpha\cdot\sigma_\alpha^2/p+$$
$$n_sn_3m_\alpha(p+n_s)\cdot\sigma_\alpha^2/p + n_sn_3p\cdot\sigma_{\epsilon_\alpha}^2,$$
$$Var(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha}) = o\big\{E^2(\widehat{\boldsymbol{S}}_{S\alpha}^T\widehat{\boldsymbol{S}}_{S\alpha})\big\},$$

$$E(\widehat{\boldsymbol{S}}_{S\beta}^T\widehat{\boldsymbol{S}}_{S\beta}) = n_2n_3m_\beta(p+n_2)\cdot\sigma_\beta^2/p + n_2n_3p\cdot\sigma_{\epsilon_\beta}^2 + 2n_2n_3n_sm_\beta\cdot\sigma_\beta^2/p+$$
$$n_sn_3m_\beta(p+n_s)\cdot\sigma_\beta^2/p + n_sn_3p\cdot\sigma_{\epsilon_\beta}^2,$$
$$Var(\widehat{\boldsymbol{S}}_{S\beta}^T\widehat{\boldsymbol{S}}_{S\beta}) = o\big\{E^2(\widehat{\boldsymbol{S}}_{S\beta}^T\widehat{\boldsymbol{S}}_{S\beta})\big\}.$$

Then Propositions A.6 and A.7 follow from Markov's inequality.

Real data analysis details

The raw MRI are downloaded from the Pediatric Imaging, Neurocognition, and Genetics (PING) study (Jernigan et al., 2016) resource. We process the MRI data locally using the standard procedures via advanced normalization tools (ANTs, (Avants et al., 2011)) to generate ROI volumes. Normalization/standardization using the ANTs software is detailed in Tustison et al. (2014) and Avants et al. (2011). We use the standard OASIS-30 Atropos template for registration and Mindboggle-101 atlases for labeling. Details of these templates and processing steps can be found in `https://mindboggle.info/data.html`, Klein and Tourville (2012) and Tustison et al. (2014). We focus on seven ROIs including thalamus proper, caudate, putamen, pallidum, hippocampus, accumbens area, and total brain volume (TBV) in this analysis. For the first six ROIs, their volumes are the mean of volumes of the corresponding left and right ROIs. For each phenotype and continuous covariate variable, we remove values greater than five times the median absolute deviation from the median value. We select subjects of European ancestry in the analysis. The age of the PING samples range from 3 to 21, with mean 12.28, and the proportion of male is 0.52.

Genotype imputation is performed on the PING dataset using standard procedures via MACH-Admix (Liu et al., 2013). A full description of the imputation procedures is detailed supplementary information of Zhao et al. (2018). We further perform the following genetic variants data quality controls on each dataset: 1) exclude subjects with more than 10% missing genotypes; 2) exclude variants with minor allele frequency less than 0.01; 3) exclude variants with larger than 10% missing genotyping rate; and 4) exclude variants that fail the Hardy-Weinberg test at 1×10^{-7} level. To obtain independent SNPs for constructing cross-trait PRS, we perform LD pruning via Plink (Purcell et al., 2007) with $R^2 = 0.2$ and window size 50. SNPs that remain after LD pruning are used in later steps as candidates for constructing PRS.

Part of the data used in the preparation of this article were obtained from the Pediatric Imaging, Neurocognition and Genetics (PING) Study database (`http://ping.chd.ucsd.`

edu/). PING was launched in 2009 by the National Institute on Drug Abuse (NIDA) and the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD) as a 2-year project of the American Recovery and Reinvestment Act. The primary goal of PING has been to create a data resource of highly standardized and carefully curated magnetic resonance imaging (MRI) data, comprehensive genotyping data, and developmental and neuropsychological assessments for a large cohort of developing children aged 3 to 20 years. The scientific aim of the project is, by openly sharing these data, to amplify the power and productivity of investigations of healthy and disordered development in children, and to increase understanding of the origins of variation in neurobehavioral phenotypes. For up-to-date information, see http://ping.chd.ucsd.edu/.

APPENDIX B: TECHNICAL DETAILS OF CHAPTER 4

Special case: independent features

Corollary B.1. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min(n, n_z, $m_{\beta\eta}$, p) $\to \infty$, for any $\omega \in (0, \infty)$, $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have*

$$A_R^2(\lambda) = A_B^2(\lambda/\omega) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\left\{1 - \lambda g(-\lambda)\right\}^2 \cdot h_\beta^2}{\left\{1 - 2\lambda g(-\lambda) + \lambda^2 \dot{g}(-\lambda)\right\} \cdot h_\beta^2 + \left\{\omega g(-\lambda) - \omega\lambda \dot{g}(-\lambda)\right\} \cdot (1 - h_\beta^2)} + o_p(1),$$

where closed-form expressions of $g(\cdot)$ and $\dot{g}(\cdot)$ can be found in equations 4.3 and 4.4. In addition,

$$\begin{aligned} A_R^2(0^+) = A_B^2(0^+) &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\left\{\frac{1+\omega-|\omega-1|}{2\omega}\right\}^2 \cdot h_\beta^2}{\frac{1+\omega-|\omega-1|}{2\omega} \cdot h_\beta^2 + \frac{\omega+1-|\omega-1|}{2|\omega-1|} \cdot (1 - h_\beta^2)} + o_p(1) \\ &= \begin{cases} h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{h_\beta^2}{h_\beta^2 + \frac{\omega}{1-\omega} \cdot (1 - h_\beta^2)} + o_p(1), & \text{if} \quad \omega < 1; \\ h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{h_\beta^2}{h_\beta^2 \cdot \omega + \frac{\omega^2}{\omega - 1} \cdot (1 - h_\beta^2)} + o_p(1), & \text{if} \quad \omega > 1. \end{cases} \end{aligned}$$

If $h_\beta^2 \in (0, 1)$, $A_R^2(\lambda)$ is maximized at λ^*, and the optimal out-of-sample R^2 is given by

$$\begin{aligned} A_R^2(\lambda^*) = A_B^2(\lambda^*/\omega) &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \left\{1 - \lambda^* g(-\lambda^*)\right\} + o_p(1) \\ &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\omega + h_\beta^2 - \sqrt{(\omega - h_\beta^2)^2 + 4\omega h_\beta^2 (1 - h_\beta^2)}}{2\omega h_\beta^2} + o_p(1). \end{aligned}$$

If $h_\beta^2 = 1$, the optimal $A_R^2(\lambda)$ is obtained as $\lambda \to 0^+$, and we have

$$\begin{aligned} A_R^2(0^+) = A_B^2(0^+) &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\omega + 1 - |\omega - 1|}{2\omega} + o_p(1) \\ &= \begin{cases} h_\eta^2 \varphi_{\beta\eta}^2 + o_p(1), & \text{if} \quad \omega < 1; \\ h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{1}{\omega} + o_p(1), & \text{if} \quad \omega > 1. \end{cases} \end{aligned}$$

Relative prediction accuracy

In this section, we study the relative prediction accuracy of marginal estimator compared to the optimal ridge estimator.

Corollary B.2. *Let* $R_R(h_\beta^2, \omega) = A_R^2(\lambda^*)/A_S^2$. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as* $min(n, n_z, m_{\beta\eta}, p) \to \infty$, *for any* $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, *and* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, *we have*

$$R_R(h_\beta^2, \omega) = \frac{(\omega + h_\beta^2)\big\{(\omega + h^2) - \sqrt{(\omega + h_\beta^2)^2 - 4\omega h_\beta^4}\big\}}{2h_\beta^4 \omega} + o_p(1) > 1 + o_p(1).$$

Moreover, for any fixed $\omega \in (0, \infty)$, $\frac{dR}{dh_\beta^2} R_B(h_\beta^2, \omega) > 0$ *on* $h_\beta^2 \in (0, 1)$; *and for any given* $h_\beta^2 \in (0, 1]$, *we have*

$$\frac{dR}{d\omega} R_R(h_\beta^2, \omega) \begin{cases} > 0, & if \quad 0 < \omega < h_\beta^2; \\ = 0, & if \quad \omega = h_\beta^2; \\ < 0, & if \quad \omega > h_\beta^2. \end{cases}$$

Corollary A.1 shows that $\widehat{\boldsymbol{\beta}}_R(\lambda^*)$ always has better asymptotic out-of-sample R^2 than $\widehat{\boldsymbol{\beta}}_S$. $R_R(h_\beta^2, \omega)$ is higher for larger h_β^2 and is not a monotone function of ω. For given h_β^2, $R_R(h_\beta^2, \omega)$ is maximized at $\omega = h_\beta^2$, with the maximum value

$$R_R^*(h_\beta^2, \omega) = \frac{2 - 2\sqrt{1 - h_\beta^2}}{h_\beta^2} + o_p(1).$$

$R_R^*(h_\beta^2, \omega)$ is an increasing function of h_β^2 on $h_\beta^2 \in (0, 1]$ and the maximun value is 2 at $h_\beta^2 = 1$. That is, for a fully heritable trait and $p = n$ in the training GWAS, we have $R_B(h_\beta^2, \omega) = A_R^2(\lambda^*)/A_S^2 = 2$. This represents the difference between n/p and $n/(n + p)$. As ω becomes large, $R_B(h_\beta^2, \omega)$ decreases, which can be viewed as a blessing of dimensionality due to the fact that the difference between n/p and $n/(n + p)$ decreases. Another interesting

question is the relative prediction accuracy between $\widehat{\boldsymbol{\beta}}_S$ and $\widehat{\boldsymbol{\beta}}_R(0^+)$, which is quantified in the following corollary.

Corollary B.3. *Let $R_{R^0}(h_\beta^2, \omega) = A_R^2(0^+)/A_S^2$. Under polygenic model (4.1) and Conditions 4.1 and 4.2, as $min(n, n_z, m_{\beta\eta}, p) \to \infty$, for any $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have*

$$R_{R^0}(h_\beta^2, \omega) = \begin{cases} \frac{h_\beta^2 + \frac{\omega}{1-\omega} \cdot (1-h_\beta^2)}{h_\beta^2 + \omega} + o_p(1), & \text{if} \quad \omega < 1; \\ \frac{h_\beta^2 \omega + (1-h_\beta^2)\frac{\omega^2}{\omega - 1}}{h_\beta^2 + \omega} + o_p(1), & \text{if} \quad \omega > 1. \end{cases}$$

It follows that for $h_\beta^2 \in (0, 1/2]$, we have

$$R_{R^0}(h_\beta^2, \omega) \begin{cases} > 1 + o_p(1), & \text{if} \quad \omega < h_\beta^2; \\ = 1 + o_p(1), & \text{if} \quad \omega = h_\beta^2; \\ < 1 + o_p(1), & \text{if} \quad \omega > h_\beta^2; \end{cases}$$

and for $h_\beta^2 \in (1/2, 1)$, we have

$$R_{R^0}(h_\beta^2, \omega) \begin{cases} > 1 + o_p(1), & \text{if} \quad \omega < h_\beta^2 \quad \text{or} \quad \omega > h_\beta^2/(2h_\beta^2 - 1); \\ = 1 + o_p(1), & \text{if} \quad \omega = h_\beta^2 \quad \text{or} \quad \omega = h_\beta^2/(2h_\beta^2 - 1); \\ < 1 + o_p(1), & \text{if} \quad h_\beta^2 < \omega < h_\beta^2/(2h_\beta^2 - 1). \end{cases}$$

And if $h_\beta^2 = 1$, then $R_{R^0}(h_\beta^2, \omega) > 1 + o_p(1)$ for any ω.

As $\widehat{\boldsymbol{\beta}}_R(0^+)$ reduces to $\widehat{\boldsymbol{\beta}}_O$ when $\omega < 1$, our results indicate that $\widehat{\boldsymbol{\beta}}_S$ can have better out-of-sample R^2 than $\widehat{\boldsymbol{\beta}}_O$ when $1 > \omega > h_\beta^2$. Thus, $\widehat{\boldsymbol{\beta}}_S$ can easily outperform $\widehat{\boldsymbol{\beta}}_O$ when h_β^2 is low. Moreover, if $h_\beta^2 \leq 0.5$, $\widehat{\boldsymbol{\beta}}_R(0^+)$ is worse than $\widehat{\boldsymbol{\beta}}_S$ for $1 < \omega$. If $h_\beta^2 > 0.5$, however, $\widehat{\boldsymbol{\beta}}_R(0^+)$ is better when ω is large.

Relative goodness-of-fit

The following corollary provides the comparison between $E_R^2(\lambda^*)$ and E_S^2, and some interesting properties of E_S^2.

Corollary B.4. *Let $Q_R(h^2, \omega) = E_R^2(\lambda^*)/E_S^2$. Under polygenic model (4.1) and Conditions 4.1 and 4.2, as $min(n, n_z, m_\beta, p) \to \infty$, for any $h_\beta^2 \in (0, 1]$, $\omega \in (0, \infty)$, and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have*

$$Q_R(h^2, \omega) > 1 + o_p(1).$$

Moreover, if $h^2 \in (0, 0.5]$, we have

$$\frac{dE}{d\omega} E_S^2(h_\beta^2, \omega) > 1 + o_p(1).$$

which indicates that $E_S^2(h_\beta^2, \omega)$ is a monotone function of ω. If $h_\beta^2 \in (0.5, 1]$, however, we have

$$\frac{dE}{d\omega} E_S^2(h_\beta^2, \omega) \begin{cases} < 0, & \text{if} \quad 0 < \omega < h_\beta^2 \cdot (2h_\beta^2 - 1); \\ = 0, & \text{if} \quad \omega = h_\beta^2 \cdot (2h_\beta^2 - 1); \\ > 0, & \text{if} \quad \omega > h_\beta^2 \cdot (2h_\beta^2 - 1), \end{cases}$$

and $E_S^2(h_\beta^2, \omega) = 4h_\beta^4/(4h_\beta^4 + 1) + o_p(1)$ at $\omega = h_\beta^2 \cdot (2h_\beta^2 - 1)$.

Mean squared prediction errors

In this section, we study the MSE of marginal estimator and illustrate the bias-variance trade-off of each estimator. We focus on the same trait prediction case in which $\boldsymbol{\beta} = \boldsymbol{\eta}$. The MSE and bias-variance decomposition (e.g., Hastie et al. (2019)) of a generic $p \times 1$ estimator

$\widehat{\boldsymbol{\beta}}$ trained on GWAS dataset $(\boldsymbol{X}, \boldsymbol{y})$ can be defined as

$$
\begin{aligned}
M^2 &= \mathrm{E}\big\{ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 | \boldsymbol{X}, \boldsymbol{y} \big\} = \mathrm{E}\big\{ \big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)^T \boldsymbol{\Sigma} \big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) | \boldsymbol{X}, \boldsymbol{y} \big\} \\
&= \big\{ \mathrm{E}\big(\widehat{\boldsymbol{\beta}} | \boldsymbol{X}, \boldsymbol{y}\big) - \boldsymbol{\beta} \big\}^T \boldsymbol{\Sigma} \big\{ \mathrm{E}\big(\widehat{\boldsymbol{\beta}} | \boldsymbol{X}, \boldsymbol{y}\big) - \boldsymbol{\beta} \big\} + \mathrm{tr}\big[\mathrm{Cov}\big(\widehat{\boldsymbol{\beta}} | \boldsymbol{X}, \boldsymbol{y}\big) \boldsymbol{\Sigma} \big] \equiv B^2 + V^2,
\end{aligned}
$$

where B^2 represents the squared bias of $\widehat{\boldsymbol{\beta}}$, and V^2 measures the total variance of the $\widehat{\boldsymbol{\beta}}$ due to the random error term. We define $M_S^2 = B_S^2 + V_S^2$, $M_R^2(\lambda) = B_R^2(\lambda) + V_R^2(\lambda)$, $M_R^2(0^+) = B_R^2(0^+) + V_R^2(0^+)$, $M_O^2 = B_O^2 + V_O^2$, and $M_B^2(\tau) = B_B^2(\tau) + V_B^2(\tau)$, for marginal, ridge, ridge-less, OLS, and BLUP estimators, respectively.

Proposition B.1. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n, n_z, m_\beta, p) \to \infty$*, for any* $h_\beta^2 \in (0, 1]$ *and* $\boldsymbol{\Sigma}$*, we have*

$$
\begin{aligned}
&M_S^2 = \Big[m\sigma_\beta^2/p \cdot \big\{ \omega b_2(\boldsymbol{\Sigma}) + b_3(\boldsymbol{\Sigma}) - 2b_2(\boldsymbol{\Sigma}) + 1 \big\} + \sigma_\epsilon^2 \cdot \omega b_2(\boldsymbol{\Sigma}) \Big] \cdot \{1 + o_p(1)\}, \\
&M_R^2(\lambda) = M_B^2(\lambda/\omega) = \Big[m\sigma_\beta^2/p \cdot \frac{v(-\lambda) - \lambda \dot{v}(-\lambda)}{v(-\lambda)^2 \omega} + \sigma_\epsilon^2 \cdot \big\{ \frac{\dot{v}(-\lambda)}{v(-\lambda)^2} - 1 \big\} \Big] \cdot \{1 + o_p(1)\}, \\
&M_R^2(0^+) = M_B^2(0^+) = \Big[m\sigma_\beta^2/p \cdot \frac{1}{v(0^+)\omega} + \sigma_\epsilon^2 \cdot \big\{ \frac{\dot{v}(0^+)}{v(0^+)^2} - 1 \big\} \Big] \cdot \{1 + o_p(1)\}, \quad \textit{and} \\
&M_O^2 = \sigma_\epsilon^2 \cdot \frac{\omega}{1 - \omega} \cdot \{1 + o_p(1)\} \quad (\omega < 1).
\end{aligned}
$$

Moreover, if $h_\beta^2 \in (0, 1)$*,* $M_R^2(\lambda)$ *is minimized at* λ^* *with the minimize value*

$$
M_R^2(\lambda^*) = M_B^2(\lambda^*/\omega) = m\sigma_\beta^2/p \cdot \Big\{ \frac{1}{v(-\lambda^*)\omega} - \frac{\lambda^*}{\omega} \Big\} \cdot \{1 + o_p(1)\}.
$$

And if $h_\beta^2 = 1$*,* $M_R^2(\lambda)$ *is minimized as* $\lambda \to 0^+$*, and thus the minimize value is* $M_R^2(0^+)$*.*

When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have the following closed-form expressions on MSE of ridge-type estimators.

Proposition B.2. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,$

$n_z, m_\beta, p) \to \infty$, *when* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, *for any* $h_\beta^2 \in (0,1)$, *we have*

$$
\begin{aligned}
&M_S^2 = \big(m\sigma_\beta^2/p \cdot \omega + \sigma_\epsilon^2 \cdot \omega\big) \cdot \{1 + o_p(1)\}, \\
&M_R^2(\lambda) = M_B^2(\lambda/\omega) = \Big[m\sigma_\beta^2/p \cdot \lambda^2 \dot{g}(-\lambda) + \sigma_\epsilon^2 \cdot \omega\big\{g(-\lambda) - \lambda\dot{g}(-\lambda)\big\}\Big] \cdot \{1 + o_p(1)\},
\end{aligned}
$$

and

$$
\begin{aligned}
M_R^2(0^+) = M_B^2(0^+) &= \Big[m\sigma_\beta^2/p \cdot \frac{(\omega-1) + |\omega - 1|}{2\omega} + \sigma_\epsilon^2 \cdot \frac{\omega + 1 - |\omega - 1|}{2|\omega - 1|}\Big] \cdot \{1 + o_p(1)\} \\
&= \begin{cases} M_O^2, & \textit{if} \quad \omega < 1; \\ \Big\{m\sigma_\beta^2/p \cdot \frac{\omega-1}{\omega} + \sigma_\epsilon^2 \cdot \frac{1}{\omega-1}\Big\} \cdot \{1 + o_p(1)\}, & \textit{if} \quad \omega > 1. \end{cases}
\end{aligned}
$$

Moreover, for $h_\beta^2 \in (0,1)$, $M_R^2(\lambda)$ *is minimized at* λ^* *with the minimize value*

$$
\begin{aligned}
M_R^2(\lambda^*) &= M_B^2(\lambda^*/\omega) = m\sigma_\beta^2 \cdot \lambda^* g(-\lambda^*) \\
&= m\sigma_\beta^2/p \cdot \frac{2\omega h_\beta^2 + \sqrt{(\omega - h_\beta^2)^2 + 4\omega h_\beta^2(1 - h_\beta^2)} - \omega - h_\beta^2}{2\omega h_\beta^2} \cdot \{1 + o_p(1)\}.
\end{aligned}
$$

For $h_\beta^2 = 1$, $M_R^2(\lambda)$ *is minimized at* $\lambda^* \to 0^+$, *and the minimize value is*

$$
M_R^2(0^+) = \begin{cases} o_p(1), & \textit{if} \quad \omega < 1; \\ m\sigma_\beta^2/p \cdot \frac{\omega-1}{\omega} \cdot \{1 + o_p(1)\}, & \textit{if} \quad \omega > 1. \end{cases}
$$

Proposition B.3. *Under the same conditions as in Proposition B.2, we have for any* ω

$$
\frac{M_S^2}{M_R^2(\lambda^*)} = \frac{1}{(1 - h_\beta^2)g(-\lambda^*)} = \frac{2\omega^2}{\sqrt{(\omega + h_\beta^2)^2 - 4h_\beta^4\omega} + h_\beta^2(2\omega - 1) - \omega} > 1 + o_p(1).
$$

Moreover, we have

$$\frac{M_S^2}{M_O^2} = \frac{1-\omega}{1-h_\beta^2} \begin{cases} > 1 + o_p(1), & \text{if} \quad 0 < \omega < h_\beta^2; \\ = 1 + o_p(1), & \text{if} \quad \omega = h_\beta^2; \\ < 1 + o_p(1), & \text{if} \quad \omega > h_\beta^2. \end{cases}$$

When $h_\beta^2 = 1$*, we have* $M_S^2/M_O^2 > 1 + o_p(1)$ *and* $M_S^2/M_B^2(0^+) = M_S^2/M_R^2(0^+) > 1 + o_p(1)$ *for any* ω*.*

Proof of Lemma 4.4

Let $\boldsymbol{\alpha}$ be a p-dimensional random vector of i.i.d. elements with mean zero and finite fourth order moment, and $\boldsymbol{A}$ be a fixed $p \times p$ matrix. Without loss of generality, we let $\mathrm{E}(\alpha_1^2) = 1$. Then, by Lemma B.26 of Bai and Silverstein (2010), for any $q \geq 1$, we have

$$\mathrm{E}\Big[\big\{\boldsymbol{\alpha}^T\boldsymbol{A}\boldsymbol{\alpha} - \mathrm{tr}(\boldsymbol{A})\big\}^q\Big] \leq c_q \cdot \Big[\big\{\mathrm{E}(\alpha_1^4)\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^T)\big\}^{q/2} + \mathrm{E}(\alpha_1^{2q})\mathrm{tr}\big\{(\boldsymbol{A}\boldsymbol{A}^T)^{q/2}\big\}\Big],$$

where c_q is some constant that only depends on q. Let $q = 2$, it follows that

$$\mathrm{Var}\Big\{\frac{\boldsymbol{\alpha}^T\boldsymbol{A}\boldsymbol{\alpha}}{\mathrm{tr}(\boldsymbol{A})}\Big\} = \mathrm{E}\Big[\big\{\frac{\boldsymbol{\alpha}^T\boldsymbol{A}\boldsymbol{\alpha}}{\mathrm{tr}(\boldsymbol{A})} - 1\big\}^2\Big] \leq c_q \cdot \Big[\frac{\mathrm{E}(\alpha_1^4)\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^T)}{\mathrm{tr}(\boldsymbol{A})^2} + \frac{\mathrm{E}(\alpha_1^4)\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^T)}{\mathrm{tr}(\boldsymbol{A})^2}\Big].$$

Let $\boldsymbol{B}_{k_1,k_2} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{k_1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{k_2}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = n_z^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$. Then, for bounded ω and any $\boldsymbol{\Sigma}$ with uniformly bounded eigenvalues, we have

$$\mathrm{tr}(\boldsymbol{B}_{k_1,k_2}) \leq p \cdot b_{k_1+k_2}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}) = O(p).$$

Then, we have

$$\frac{\mathrm{tr}\big(\boldsymbol{B}_{k_1,k_2}\boldsymbol{B}_{k_1,k_2}^T\big)}{\big\{\mathrm{tr}(\boldsymbol{B}_{k_1,k_2})\big\}^2} = \frac{\mathrm{tr}(\boldsymbol{B}_{2k_1,2k_2})}{\big\{\mathrm{tr}(\boldsymbol{B}_{k_1,k_2})\big\}^2} = O(\frac{p}{p^2}) = O(\frac{1}{p}) = o(1).$$

It follows that

$$\mathrm{Var}\Big\{\frac{\boldsymbol{\alpha}^T\boldsymbol{B}_{k_1,k_2}\boldsymbol{\alpha}}{\mathrm{tr}(\boldsymbol{B}_{k_1,k_2})}\Big\} = \mathrm{E}\Big[\Big\{\frac{\boldsymbol{\alpha}^T\boldsymbol{B}_{k_1,k_2}\boldsymbol{\alpha}}{\sigma_\alpha^2\mathrm{tr}(\boldsymbol{B}_{k_1,k_2})} - 1\Big\}^2\Big] = o(1).$$

Thus, by Markov's inequality, we have

$$\boldsymbol{\alpha}^T\boldsymbol{B}_{k_1,k_2}\boldsymbol{\alpha} = \sigma_\alpha^2 \cdot \mathrm{tr}(\boldsymbol{B}_{k_1,k_2}) \cdot \{1 + o_p(1)\}.$$

Useful trace results for ridge-type estimators

Here we summarize some results that are used frequently in our analysis of ridge-type estimators, which are based on Lemma 4.2.

Lemma B.1. *Under Condition 4.1, as* $min(n,p) \to \infty$, *for any* $\lambda > 0$, *we have*

$$tr\big\{\boldsymbol{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\big\} \to_{a.s.} \frac{1}{\lambda v(-\lambda)} - 1, \quad \textit{and}$$
$$tr\big\{\boldsymbol{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-2}\big\} \to_{a.s.} \frac{1}{n} \cdot \frac{v(-\lambda) - \lambda\dot{v}(-\lambda)}{(\lambda v(-\lambda))^2}.$$

When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, *we have closed-form limits*

$$tr\big\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\big\} \to_{a.s.} \omega g(-\lambda) = \frac{\sqrt{(1-\omega+\lambda)^2 + 4\omega\lambda} - (1-\omega+\lambda)}{2\lambda},$$
$$\textit{and} \quad tr\big\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-2}\big\} \to_{a.s.} \frac{p}{n^2} \cdot \dot{g}(-\lambda) = \frac{p}{n^2} \cdot \frac{(\omega-1) + \frac{(\omega+1)\lambda+(\omega-1)^2}{\sqrt{(1-\omega+\lambda)^2+4\omega\lambda}}}{2\omega\lambda^2}.$$

Moreover, at the optimal $\lambda^* = \omega \cdot (1-h_\beta^2)/h_\beta^2$*, we have*

$$\begin{aligned}
A_R^2(\lambda^*) &= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \{1 - \lambda^* g(-\lambda^*)\} + o_p(1) \\
&= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\omega + h_\beta^2 - \sqrt{(h_\beta^2 - 2\omega h_\beta^2 + \omega)^2 + 4\omega^2 h_\beta^2 (1-h_\beta^2)}}{2\omega h_\beta^2} + o_p(1) \\
&= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\omega + h_\beta^2 - \sqrt{(\omega + h_\beta^2)^2 - 4\omega h_\beta^4}}{2\omega h_\beta^2} + o_p(1) \\
&= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\omega + h_\beta^2 - \sqrt{(\omega - h_\beta^2)^2 + 4\omega h_\beta^2 (1-h_\beta^2)}}{2\omega h_\beta^2} + o_p(1).
\end{aligned}$$

Lemma B.2. *Under Condition 4.1, when* $\boldsymbol{\Sigma} = \boldsymbol{I}_p$*, as* $min(n,p) \to \infty$*,* $\lambda \to 0^+$*, we have the following closed-form limits*

$$\begin{gathered}
g(-\lambda) \to_{a.s.} \frac{\frac{1+\omega}{|1-\omega|} - 1}{2\omega}, \quad \lambda g(-\lambda) \to_{a.s.} \frac{|1-\omega| - (1-\omega)}{2\omega}, \\
\lambda \dot{g}(-\lambda) \to_{a.s.} \frac{2\lambda}{\{(1-\omega+\lambda)^2 + 4\omega\lambda\}^{3/2}} = 0, \quad \textit{and} \\
\lambda^2 \dot{g}(-\lambda) \to_{a.s.} \frac{(\omega-1) + |1-\omega|}{2\omega}.
\end{gathered}$$

The following lemma is used to show the equivalence of prediction accuracy of ridge estimator and BLUP, which can be easily proved by applying singular value decomposition on $\boldsymbol{X}$.

Lemma B.3. *Under Condition 4.1, as* $min(n,p) \to \infty$*, for any* $\lambda > 0$ *and arbitrary* $\boldsymbol{\Sigma}$*, we have*

$$\begin{gathered}
tr\big\{(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}_n)^{-1} \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T\big\} = tr\big\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}\big\}, \quad \textit{and} \\
tr\big\{(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}_n)^{-2} \boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^T\big\} = tr\big\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-2} \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}\big\}.
\end{gathered}$$

Proofs of marginal estimator: out-of-sample

Proposition B.4. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ m_{\beta\eta},\ p) \to \infty$*, for any* $\omega \in (0, \infty)$*,* $h_\beta^2, h_\eta^2 \in (0, 1]$*,* $\varphi_{\beta\eta} \in [-1, 1]$*, and* $\boldsymbol{\Sigma}$*, we have*

$$\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)}{n_z m_\eta \cdot \sigma_\eta^2/p + n_z \cdot \sigma_{\epsilon_z}^2} = 1 + o_p(1),$$
$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{nn_z m_\beta \cdot \{(n+1)b_3(\boldsymbol{\Sigma}) + pb_2(\boldsymbol{\Sigma})\} \cdot \sigma_\beta^2/p + nn_z p b_2(\boldsymbol{\Sigma}) \cdot \sigma_\epsilon^2} = 1 + o_p(1),$$

and

$$\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)^T\boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{nn_z m_{\beta\eta} b_2(\boldsymbol{\Sigma}) \cdot \sigma_{\beta\eta}/p} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$A_S^2 = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \Big\{\frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \frac{\omega}{h_\beta^2} \cdot \frac{1}{b_2(\boldsymbol{\Sigma})}\Big\}^{-1} + o_p(1).$$

Proofs of marginal estimator: meta-analysis

Under polygenic model (4.1) and Conditions 4.1 and 4.2, suppose we have GWAS $(\boldsymbol{X}_i, \boldsymbol{y}_i)$, with sample sizes $(n_i, \cdots, n_k)$ and p SNPs, $i = 1, \cdots, k$, $k \in (0, \infty)$, let $\widehat{\boldsymbol{B}} = [\widehat{\boldsymbol{\beta}}_1^T, \cdots, \widehat{\boldsymbol{\beta}}_k^T]$ be the $p \times k$ matrix of marginal estimators from the k GWAS. Let $\boldsymbol{d} = (d_i, \cdots, d_k)^T$ be an $k \times 1$ vector of weights, and let $\widehat{\boldsymbol{B}}(\boldsymbol{d}) = \widehat{\boldsymbol{B}}\boldsymbol{d}$ be the aggregated summary statistics. As $\min(n_1, \cdots, n_k,\ n_z,\ m_{\beta\eta},\ p) \to \infty$, for any $\omega \in (0, \infty)$, $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, $\boldsymbol{d}$, and

$\boldsymbol{\Sigma}$, we have

$$A_S^2(\boldsymbol{d}) = \Big\{ \frac{\big(\sum_{i=1}^k d_i \widehat{\boldsymbol{\beta}}_i\big)^T \boldsymbol{Z}^T \big(\boldsymbol{Z}\eta + \boldsymbol{\epsilon}_z\big)}{\|\boldsymbol{Z}\eta + \boldsymbol{\epsilon}_z\| \|\boldsymbol{Z}\widehat{B}\boldsymbol{d}\|} \Big\}^2$$
$$= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\big(\sum_{i=1}^k d_i\big)^2 \cdot b_2(\boldsymbol{\Sigma})^2}{\big(\sum_{i=1}^k d_i \widehat{\boldsymbol{\beta}}_i\big)^T \boldsymbol{Z}^T \boldsymbol{Z} \big(\sum_{i=1}^k d_i \widehat{\boldsymbol{\beta}}_i\big)} + o_p(1).$$

Note that for $i \neq j, (i,j) \in (1, \cdots, k)$, we have

$$\big(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_i\big)^T \big(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_j\big) = \frac{1}{n_i n_j} \mathrm{E}\big(\boldsymbol{\beta}^T \boldsymbol{X}_i^T \boldsymbol{X}_i \boldsymbol{Z}^T \boldsymbol{Z} \boldsymbol{X}_j^T \boldsymbol{X}_j \boldsymbol{\beta}\big) \cdot \{1 + o_p(1)\}$$
$$= m_\beta n_z \cdot b_3(\boldsymbol{\Sigma}) \cdot \sigma_\beta^2 / p \cdot \{1 + o_p(1)\}$$

and

$$\big(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_i\big)^T \big(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_i\big) = \frac{1}{n_i^2} \mathrm{E}\big(\boldsymbol{\beta}^T \boldsymbol{X}_i^T \boldsymbol{X}_i \boldsymbol{Z}^T \boldsymbol{Z} \boldsymbol{X}_i^T \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i^T \boldsymbol{X}_i^T \boldsymbol{Z}^T \boldsymbol{Z} \boldsymbol{X}_i^T \boldsymbol{\epsilon}_i\big) \cdot \{1 + o_p(1)\}$$
$$= m_\beta n_z \cdot \Big\{ b_3(\boldsymbol{\Sigma}) + \frac{p}{n_i h_\beta^2} b_2(\boldsymbol{\Sigma}) \Big\} \cdot \sigma_\beta^2 / p \cdot \{1 + o_p(1)\}.$$

It follows that

$$A_S^2(\boldsymbol{d}) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\big(\sum_{i=1}^k d_i\big)^2 \cdot b_2(\boldsymbol{\Sigma})^2}{\sum_{i=1}^k d_i^2 \cdot \{b_3(\boldsymbol{\Sigma}) + \frac{p}{n_i h_\beta^2} b_2(\boldsymbol{\Sigma})\} + 2 \sum_{i \neq j}^{(i,j) \in (1,\cdots,k)} d_i d_j \cdot b_3(\boldsymbol{\Sigma})} + o_p(1)$$
$$= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\big(\sum_{i=1}^k d_i\big)^2 \cdot b_2(\boldsymbol{\Sigma})^2}{\big(\sum_{i=1}^k d_i\big)^2 \cdot b_3(\boldsymbol{\Sigma}) + \big(\sum_{i=1}^k d_i^2 \frac{p}{n_i h_\beta^2}\big) \cdot b_2(\boldsymbol{\Sigma})} + o_p(1)$$
$$= h_\eta^2 \varphi_{\beta\eta}^2 \cdot \Big\{ \frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \frac{\sum_{i=1}^k d_i^2 / n_i}{(\sum_{i=1}^k d_i)^2} \cdot \frac{p}{b_2(\boldsymbol{\Sigma}) h_\beta^2} \Big\}^{-1} + o_p(1).$$

Therefore, when $d_i = n_i$, we have

$$A_S^2(\boldsymbol{a}^*) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \Big\{ \frac{b_3(\boldsymbol{\Sigma})}{b_2(\boldsymbol{\Sigma})^2} + \frac{p}{\sum_{i=1}^k n_i} \cdot \frac{1}{b_2(\boldsymbol{\Sigma}) h_\beta^2} \Big\}^{-1} + o_p(1).$$

Proofs of marginal estimator: in-sample

Proposition B.5. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ m_\beta,\ p) \to \infty$*, for any* $\omega \in (0, \infty)$*,* $h_\beta^2 \in (0, 1]$*, and* $\boldsymbol{\Sigma}$*, we have*

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{nm_\beta \cdot \sigma_\beta^2/p + n \cdot \sigma_\epsilon^2} = 1 + o_p(1),$$
$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{n^3m_\beta \cdot \{b_3(\boldsymbol{\Sigma}) + 3\omega b_2(\boldsymbol{\Sigma}) + \omega^2\} \cdot \sigma_\beta^2/p + n^2p \cdot \{b_2(\boldsymbol{\Sigma}) + \omega\} \cdot \sigma_\epsilon^2} = 1 + o_p(1),$$

and

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T\boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{n^2m_\beta\{b_2(\boldsymbol{\Sigma}) + \omega\} \cdot \sigma_\beta^2/p + np \cdot \sigma_\epsilon^2} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$E_S^2 = \frac{\{b_2(\boldsymbol{\Sigma})h_\beta^2 + \omega\}^2}{\{b_2(\boldsymbol{\Sigma})h_\beta^2 + \omega\}^2 + b_2(\boldsymbol{\Sigma})\omega + \{b_3(\boldsymbol{\Sigma}) - b_2(\boldsymbol{\Sigma})^2h_\beta^2\}h_\beta^2} + o_p(1).$$

For the special case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have $b_3(\boldsymbol{\Sigma}) = b_2(\boldsymbol{\Sigma}) = b_1(\boldsymbol{\Sigma}) = 1$ in the above propositions.

Proofs of ridge-type estimators

Proposition B.6. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ m_{\beta\eta},\ p) \to \infty$*, for any* $\omega \in (0, \infty)$*,* $h_\beta^2, h_\eta^2 \in (0, 1]$*,* $\varphi_{\beta\eta} \in [-1, 1]$*, and* $\boldsymbol{\Sigma}$*, we have*

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{n_zm_\beta \cdot \sigma_\beta^2/p + n_z\omega/(1-\omega) \cdot \sigma_\epsilon^2} = 1 + o_p(1),$$

and

$$\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)^T\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{n_zm_{\beta\eta} \cdot \sigma_{\beta\eta}/p} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$A_O^2 = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \Big\{1 + \frac{1-h_\beta^2}{h_\beta^2} \cdot \frac{\omega}{1-\omega}\Big\}^{-1} + o_p(1). \quad (\omega < 1)$$

Proposition B.7. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min(n, n_z, m_β, p) $\to \infty$, for any $\omega \in (0, \infty)$, $h_\beta^2 \in (0, 1]$, and $\mathbf{\Sigma}$, we have*

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{nm_\beta \cdot \sigma_\beta^2/p + p \cdot \sigma_\epsilon^2} = 1 + o_p(1),$$

and

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{nm_\beta \cdot \sigma_\beta^2/p + p \cdot \sigma_\epsilon^2} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$E_O^2 = \big\{h_\beta^2 + \omega(1-h_\beta^2)\big\} \cdot \{1 + o_p(1)\}. \quad (\omega < 1)$$

Proposition B.8. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min(n, n_z, $m_{\beta\eta}$, p) $\to \infty$, for any $\omega \in (0, \infty)$, $h_\beta^2, h_\eta^2 \in (0, 1]$, $\varphi_{\beta\eta} \in [-1, 1]$, and $\mathbf{\Sigma}$, we have*

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{V_{R1} + V_{R2}} = 1 + o_p(1),$$

where

$$\begin{aligned} V_{R1} &= n_z m_\beta \cdot \Big[1 - \frac{2\lambda}{\omega} tr\{\mathbf{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\} + \frac{\lambda^2 n}{\omega} \cdot tr\{\mathbf{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-2}\}\Big] \cdot \sigma_\beta^2/p \\ &= n_z m_\beta \cdot \Big[1 - \frac{2\lambda}{\omega}\Big\{\frac{1}{\lambda v(-\lambda)} - 1\Big\} + \frac{\lambda^2}{\omega} \cdot \frac{v(-\lambda) - \lambda\dot{v}(-\lambda)}{\{\lambda v(-\lambda)\}^2}\Big] \cdot \sigma_\beta^2/p, \end{aligned}$$

and

$$V_{R2} = n_z \cdot \Big[tr\{\boldsymbol{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\} + \lambda n \cdot tr\{\boldsymbol{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-2}\}] \cdot \sigma_\epsilon^2$$
$$= n_z \cdot \Big[\Big\{ \frac{1}{\lambda v(-\lambda)} - 1 \Big\} - \lambda \cdot \frac{v(-\lambda) - \lambda \dot{v}(-\lambda)}{\{\lambda v(-\lambda)\}^2} \Big] \cdot \sigma_\epsilon^2.$$

In addition, we have

$$\frac{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)^T \boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{C_{R1}} = 1 + o_p(1),$$

where

$$C_{R1} = n_z m_{\beta\eta} \cdot [1 - \frac{\lambda}{\omega} \cdot tr\{\boldsymbol{\Sigma}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\}] \cdot \sigma_{\beta\eta}/p$$
$$= n_z m_{\beta\eta} \cdot [1 - \frac{\lambda}{\omega} \cdot \{\frac{1}{\lambda v(-\lambda)} - 1\}] \cdot \sigma_{\beta\eta}/p.$$

By continuous mapping theorem, we have

$$A_R^2(\lambda) = h_\eta^2 \varphi_{\beta\eta}^2 \cdot \frac{\big[1 + \frac{\lambda}{\omega}\{1 - \frac{1}{\lambda v(-\lambda)}\}\big]^2 \cdot h_\beta^2}{h_\beta^2 \cdot \Big[1 + \frac{\lambda}{\omega}\big\{2 - \frac{1}{\lambda v(-\lambda)} - \frac{\dot{v}(-\lambda)}{v(-\lambda)^2}\big\}\Big] + (1 - h_\beta^2) \cdot \big\{\frac{\dot{v}(-\lambda)}{v(-\lambda)^2} - 1\big\}} + o_p(1).$$

Similar to Theorem 2.1 of Dobriban and Wager (2018), $A_R^2(\lambda)$ is optimized at $\lambda^* = \omega \cdot (1 - h_\beta^2)/h_\beta^2$, where the second order term $\dot{v}(-\lambda)$ disappears. In Theorem 2.1 of Dobriban and Wager (2018), they set $\gamma = p/n$, and the signal to noise ratio to α^2, thus, their optimal λ is $\lambda^* = \gamma\alpha^{-2}$. The $A_R^2(0^+)$ can be obtained by taking $\lambda \to 0^+$, with careful exchanging limits as $n, p \to \infty$ and $\lambda \to 0^+$, detailed in Threorem 4 of Hastie et al. (2019). When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, using the results in Lemmas B.1 and B.2, we have closed-form expressions for $A_R^2(\lambda)$, $A_R^2(\lambda^*)$, and $A_R^2(0^+)$.

Proposition B.9. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min(n, n_z,*

$m_\beta,\, p) \to \infty$, *for any* $\omega \in (0, \infty)$, $h_\beta^2 \in (0, 1]$ *and* $\boldsymbol{\Sigma}$, *we have*

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{V_{R3} + V_{R4}} = 1 + o_p(1),$$

where

$$\begin{aligned} V_{R3} &= nm_\beta \cdot \Big[1 - \frac{2\lambda}{p} \cdot tr\big\{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\big\} + \\ &\qquad \frac{\lambda^2}{\omega} \cdot tr\big\{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-2}\big\}\Big] \cdot \sigma_\beta^2/p \\ &= nm_\beta \cdot \big\{1 - 2\lambda + 3\lambda^2 g(-\lambda) - \lambda^3 \dot{g}(-\lambda)\big\} \cdot \sigma_\beta^2/p, \end{aligned}$$

and

$$\begin{aligned} V_{R4} &= p \cdot \Big[1 - \frac{2\lambda}{\omega} \cdot tr\big\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\big\} + \frac{\lambda^2 n}{\omega} \cdot tr\big\{(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-2}\big\}\Big] \cdot \sigma_\epsilon^2 \\ &= p \cdot \{1 - 2\lambda g(-\lambda) + \lambda^2 \dot{g}(-\lambda)\} \cdot \sigma_\epsilon^2. \end{aligned}$$

In addition, we have

$$\frac{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)}{C_{R2}} = 1 + o_p(1),$$

where

$$\begin{aligned} C_{R2} &= nm_\beta \cdot \Big[1 - \frac{\lambda}{p} \cdot tr\big\{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\big\}\Big] \cdot \sigma_\beta^2/p + \\ &\qquad tr\big\{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\big\} \cdot \sigma_\epsilon^2 \\ &= nm_\beta \cdot \{1 - \lambda + \lambda^2 g(-\lambda)\} \cdot \sigma_\beta^2/p + p \cdot \{1 - \lambda g(-\lambda)\} \cdot \sigma_\epsilon^2. \end{aligned}$$

By continuous mapping theorem, we have

$$E_R^2(\lambda) = \frac{\Big[h_\beta^2 \cdot \big\{1-\lambda+\lambda^2 g(-\lambda)\big\} + (1-h_\beta^2)\cdot\omega\big\{1-\lambda g(-\lambda)\big\}\Big]^2}{h_\beta^2 \cdot \big\{1-2\lambda+3\lambda^2 g(-\lambda)-\lambda^3\dot{g}(-\lambda)\big\} + (1-h_\beta^2)\cdot\omega\big\{1-2\lambda+\lambda^2\dot{g}(-\lambda)\big\}} + o_p(1).$$

Different from $A_R^2(\lambda)$, $E_R^2(\lambda)$ is minimized as $\lambda \to 0$, which means the over-fitting of the model in high-dimensions.

Intermediate results

Proposition B.10. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ m_{\beta\eta},\ p) \to \infty$*, for any* $\omega \in (0,\infty)$*,* $h_\beta^2, h_\eta^2 \in (0,1]$*,* $\varphi_{\beta\eta} \in [-1,1]$*, and* $\mathbf{\Sigma}$*, we have*

$$\begin{aligned}
&E\Big\{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)}+\boldsymbol{\epsilon}_z\big)^T\boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)\Big\} = nn_z m_{\beta\eta} b_2(\mathbf{\Sigma})\cdot\sigma_{\beta\eta}/p,\\
&E\Big\{\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)}+\boldsymbol{\epsilon}_z\big)^T\big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)}+\boldsymbol{\epsilon}_z\big)\Big\} = n_z m_\eta\cdot\sigma_\eta^2/p + n_z\cdot\sigma_{\epsilon_z}^2,\\
&E\Big\{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)^T\boldsymbol{X}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)\Big\}\\
&= nn_z m_\beta\cdot\{(n+1)b_3(\mathbf{\Sigma})+pb_2(\mathbf{\Sigma})\}\cdot\sigma_\beta^2/p + nn_z p b_2(\mathbf{\Sigma})\cdot\sigma_\epsilon^2,
\end{aligned}$$

$$\begin{aligned}
&E\Big\{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)^T\boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)\Big\} = n^2 m_\beta\{b_2(\mathbf{\Sigma})+\omega\}\cdot\sigma_\beta^2/p + np\cdot\sigma_\epsilon^2,\\
&E\Big\{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)\Big\} = nm_\beta\cdot\sigma_\beta^2/p + n\cdot\sigma_\epsilon^2, \quad \textit{and}\\
&E\Big\{\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)}+\boldsymbol{\epsilon}\big)\Big\}\\
&= n^3 m_\beta\cdot\{b_3(\mathbf{\Sigma})+3\omega b_2(\mathbf{\Sigma})+\omega^2\}\cdot\sigma_\beta^2/p + n^2 p\cdot\{b_2(\mathbf{\Sigma})+\omega\}\cdot\sigma_\epsilon^2.
\end{aligned}$$

Then Propositions B.4 and B.5 follow from Lemma 4.4, the concentration of marginal estimator quadratic forms. We note that the uniform boundness of $\lambda_i(\mathbf{\Sigma})$ is not a necessary condition to have the asymptotic limits of A_S^2 and E_S^2. Given boundness condition on high order moments of $H(t)$, i.e., $b_6(\mathbf{\Sigma}) < \infty$, both A_S^2 and E_S^2 can have asymptotic limits by Markov's inequality. However, the uniform boundness of $\lambda_i(\mathbf{\Sigma})$ in Condition 1 is required for

ridge-less estimator in which $\lambda \to 0^+$, see Hastie et al. (2019) for more details.

Proposition B.11. *Under polygenic model (4.1) and Conditions 4.1 and 4.2, as min*$(n,\ n_z,\ m_{\beta\eta},\ p) \to \infty$, *for any* $\omega \in (0,\infty)$, $h_\beta^2, h_\eta^2 \in (0,1]$, $\varphi_{\beta\eta} \in [-1,1]$, *and* $\boldsymbol{\Sigma}$, *we have almost surely that*

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)^T \boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)\Big\} = n_z m_{\beta\eta}\cdot \sigma_{\beta\eta}/p,\\
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)\Big\}\\
&= n_z m_\beta\cdot\sigma_\beta^2/p + n_z\omega/(1-\omega)\cdot\sigma_\epsilon^2,
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)\Big\} = n m_\beta\cdot\sigma_\beta^2/p + p\cdot\sigma_\epsilon^2,\\
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)\Big\}\\
&= n m_\beta\cdot\sigma_\beta^2/p + p\cdot\sigma_\epsilon^2,
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{Z}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_z\big)^T \boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)\Big\}\\
&= n_z m_{\beta\eta}\cdot[1 - \frac{\lambda}{\omega}\cdot\{\frac{1}{\lambda v(-\lambda)} - 1\}]\cdot\sigma_{\beta\eta}/p,\\
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)\Big\}\\
&= n_z m_\beta\cdot\Big[1 - \frac{2\lambda}{\omega}\Big\{\frac{1}{\lambda v(-\lambda)} - 1\Big\} + \frac{\lambda^2}{\omega}\cdot\frac{v(-\lambda) - \lambda\dot{v}(-\lambda)}{\{\lambda v(-\lambda)\}^2}\Big]\cdot\sigma_\beta^2/p +\\
&\qquad n_z\cdot\Big[\Big\{\frac{1}{\lambda v(-\lambda)} - 1\Big\} - \lambda\cdot\frac{v(-\lambda) - \lambda\dot{v}(-\lambda)}{\{\lambda v(-\lambda)\}^2}\Big]\cdot\sigma_\epsilon^2,
\end{aligned}
$$

$$
\begin{aligned}
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big) \Big\} \\
&= nm_\beta \cdot \{1 - \lambda + \lambda^2 g(-\lambda)\} \cdot \sigma_\beta^2/p + p \cdot \{1 - \lambda g(-\lambda)\} \cdot \sigma_\epsilon^2, \quad \textit{and} \\
&E\Big\{ \big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big)^T \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\big(\boldsymbol{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}\big) \Big\} \\
&= nm_\beta \cdot \big\{1 - 2\lambda + 3\lambda^2 g(-\lambda) - \lambda^3 \dot{g}(-\lambda)\big\} \cdot \sigma_\beta^2/p + p \cdot \{1 - 2\lambda g(-\lambda) + \lambda^2 \dot{g}(-\lambda)\} \cdot \sigma_\epsilon^2.
\end{aligned}
$$

These results are based on Lemmas B.1 and B.2. Then Propositions B.6 - B.9 follow from the concentration of ridge-type quadratic forms.

Real data analysis details

The raw MRI, covariates and genetic data are downloaded from each data resource. We process the MRI data locally using consistent procedures via advanced normalization tools (ANTs, Avants et al. (2011)) to generate ROI volumes for each dataset. Normalization steps using the ANTs software are detailed in Tustison et al. (2014) and Avants et al. (2011). We use the standard OASIS-30 Atropos template for registration and Mindboggle-101 atlases for labeling. Details of these templates and processing steps can be found in `https://mindboggle.info/data.html`, Klein and Tourville (2012) and Tustison et al. (2014). For each phenotype and continuous covariate variable, we further remove values greater than five times the median absolute deviation from the median value. We use imputed SNP data in real data analysis. We perform the following genetic variants data quality controls on each dataset: 1) exclude subjects with more than 10% missing genotypes; 2) exclude variants with minor allele frequency less than 0.01; 3) exclude variants with larger than 10% missing genotyping rate; 4) exclude variants that fail the Hardy-Weinberg test at 1×10^{-7} level; and 5) remove variants with imputation INFO score less than 0.8. All individuals are aged between 3 and 92 years. More cohort information of these studies can be found in Zhao et al. (2019).

Part of data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. HCP data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and by the McDonnell Center for Systems Neuroscience at Washington University.

Part of the data used in the preparation of this article were obtained from the Pediatric Imaging, Neurocognition and Genetics (PING) Study database (`http://ping.chd.ucsd.edu/`). PING was launched in 2009 by the National Institute on Drug Abuse (NIDA) and the Eunice Kennedy Shriver National Institute Of Child Health & Human Development

(NICHD) as a 2-year project of the American Recovery and Reinvestment Act. The primary goal of PING has been to create a data resource of highly standardized and carefully curated magnetic resonance imaging (MRI) data, comprehensive genotyping data, and developmental and neuropsychological assessments for a large cohort of developing children aged 3 to 20 years. The scientific aim of the project is, by openly sharing these data, to amplify the power and productivity of investigations of healthy and disordered development in children, and to increase understanding of the origins of variation in neurobehavioral phenotypes. For up-to-date information, see `http://ping.chd.ucsd.edu/`.

REFERENCES

1000-Genomes-Project-Consortium. (2015). A global reference for human genetic variation. *Nature* **526,** 68–74.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage* **54,** 2033–2044.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature Communications* **9,** 1825.

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32,** 283–285.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36,** 199–227.

Biton, A., Traut, N., Poline, J.-B., Aribisala, B. S., Bastin, M. E., Bülow, R., Cox, S. R., Deary, I. J., Grabe, H. J., Hagenaars, S., Hashimoto, R., Maniega, S. M., Nauck, M., Royle, N. A., Teumer, A., Hernandez, M. V., Völker, U., Wardlaw, J. M., Wittfeld, K., , Bourgeron, T., and Toro, R. (2019). Polygenic architecture of human neuroanatomical diversity. *bioRxiv 592337* .

Bogdan, R., Baranger, D. A., and Agrawal, A. (2018). Polygenic risk scores in clinical psychology: bridging genomic risk to individual differences. *Annual Review of Clinical Psychology* **14,** 119–157.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169,** 1177–1186.

Brücke, C., Bock, A., Huebl, J., Krauss, J. K., Schönecker, T., Schneider, G.-H., Brown, P., and Kühn, A. A. (2013). Thalamic gamma oscillations correlate with reaction time in a go/nogo task in patients with essential tremor. *Neuroimage* **75,** 36–45.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47,** 1236–1241.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., of the Psychiatric Genomics Consortium, S. W. G., et al. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47,** 291–295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209.

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics* **10,** 1–59.

Caldiroli, A., Buoli, M., van Haren, N. E., de Nijs, J., Altamura, A. C., and Cahn, W. (2018). The relationship of iq and emotional processing with insula volume in schizophrenia. *Schizophrenia Research* **202,** 141–148.

Chen, C.-H., Peng, Q., Schork, A. J., Lo, M.-T., Fan, C.-C., Wang, Y., Desikan, R. S., Bettella, F., Hagler, D. J., McCabe, C., et al. (2015). Large-scale genomics unveil polygenic architecture of human cortical surface area. *Nature Communications* **6,** 7549.

Choi, S. W., Heng Mak, T. S., and O'Reilly, P. F. (2018). A guide to performing polygenic risk score analyses. *bioRxiv 416545* .

Clarke, T., Lupton, M., Fernandez-Pujals, A., Starr, J., Davies, G., Cox, S., Pattie, A., Liewald, D., Hall, L., MacIntyre, D., et al. (2016). Common polygenic risk for autism spectrum disorder (asd) is associated with cognitive ability in the general population. *Molecular Psychiatry* **21,** 419–425.

Consortium, I. H. . et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20,** 273–297.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* **185,** 1021–1031.

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3,** e3395.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nature Genetics* **48,** 1284–1287.

Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., Hagenaars, S. P., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature Communications* **9,** 2098.

Davies, G., Marioni, R., Liewald, D., Hill, W., Hagenaars, S., Harris, S., Ritchie, S., Luciano, M., Fawns-Ritchie, C., Lyall, D., et al. (2016). Genome-wide association study of cognitive functions and educational attainment in uk biobank (n= 112 151). *Molecular Psychiatry* **21,** 758–767.

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genetics* **9,** e1003608.

Dicker, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electronic Journal of Statistics* **7,** 1806–1834.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22,** 1–37.

Dicker, L. H. and Erdogdu, M. A. (2017). Flexible results for quadratic forms with applications to variance components estimation. *The Annals of Statistics* **45,** 386–414.

Dobriban, E. and Sheng, Y. (2018). Distributed linear regression by averaging. *arXiv preprint arXiv:1810.00412* .

Dobriban, E. and Sheng, Y. (2019). One-shot distributed ridge regression in high dimensions. *arXiv preprint arXiv:1903.09321* .

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46,** 247–279.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* **9,** e1003348.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445* .

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields* **170,** 95–175.

Evans, L., Tahmasbi, R., Vrieze, S., Abecasis, G., Das, S., Gazal, S., Bjelland, D., Goddard, M., Neale, B., Yang, J., et al. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics* **50,** 737–745.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74,** 37–65.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 849–911.

Fan, J., Shao, Q.-M., and Zhou, W.-X. (2018). Are discoveries spurious? distributions of maximum spurious correlations and their applications. *The Annals of Statistics* **46,** 989–1017.

Feng, L. and Zhang, C.-H. (2017). Sorted concave penalized regression. *arXiv preprint arXiv:1712.09941* .

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47,** 1091–1098.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *bioRxiv 416859* .

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136,** 245–257.

Guo, X. and Cheng, G. (2018). Moderate-dimensional inferences on quadratic functionals in ordinary least squares. *arXiv preprint arXiv:1810.01323* .

Guo, Z., Wang, W., Cai, T. T., and Li, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association* **114,** 358–369.

Gurdasani, D., Barroso, I., Zeggini, E., and Sandhu, M. S. (2019). Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics* **20,** 520–535.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48,** 245–252.

Hagenaars, S. P., Harris, S. E., Davies, G., Hill, W. D., Liewald, D. C., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., Cullen, B., Malik, R., et al. (2016). Shared genetic aetiology between cognitive functions and physical and mental health in uk biobank (n= 112 151) and 24 gwas consortia. *Molecular Psychiatry* **21,** 1624–1632.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560* .

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* **21,** 309–310.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31,** 423–447.

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., et al. (2015). Common genetic variants influence human subcortical brain structures. *Nature* **520,** 224–229.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Holmes, J. B., Speed, D., and Balding, D. J. (2019). Summary statistic analyses can mistake confounding bias for heritability. *bioRxiv 532069* .

Hsu, D., Kakade, S. M., and Zhang, T. (2011). Random design analysis of ridge regression. *arXiv preprint arXiv:1106.2363* .

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S., Yu, Z., Li, B., Gu, J., Muchnik, S., et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics* **51,** 568–576.

Jansen, P. R., Nagel, M., Watanabe, K., Wei, Y., Savage, J. E., de Leeuw, C. A., van den Heuvel, M. P., van der Sluis, S., and Posthuma, D. (2019). Gwas of brain volume on 54,407 individuals and cross-trait analysis with intelligence identifies shared genomic loci and genes. *bioRxiv 613489* .

Jernigan, T. L., Brown, T. T., Hagler Jr, D. J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., et al. (2016). The pediatric imaging, neurocognition, and genetics (ping) data repository. *Neuroimage* **124,** 1149–1154.

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics* **44,** 2127–2160.

Jiang, L., Zheng, Z., Qi, T., Kemper, K., Wray, N., Visscher, P., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51,** 1749–1755.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50,** 1219–1224.

Klein, A. and Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience* **6,** 171.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields* **151,** 233–264.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88,** 365–411.

Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under stein's loss. *Bernoulli* **24,** 3791–3832.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Linnér, R. K., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics* **50,** 1112–1121.

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28,** 2540–2542.

Li, C., Yang, C., Gelernter, J., and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics* **5,** 639–650.

Liu, E. Y., Li, M., Wang, W., and Li, Y. (2013). Mach-admix: genotype imputation for admixed populations. *Genetic Epidemiology* **37,** 25–37.

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics* **47,** 1385–1392.

Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R. L., Jiang, T., Hu, Y., Chang, D., Jin, C., Dai, W., et al. (2017). A powerful approach to estimating annotation-stratified genetic covariance via gwas summary statistics. *The American Journal of Human Genetics* **101,** 939–964.

Ma, R. and Dicker, L. H. (2019). The mahalanobis kernel for heritability estimation in genome-wide association studies: fixed-effects and random-effects methods. *arXiv preprint arXiv:1901.02936* .

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik* **114,** 507–536.

Martin, A. R., Daly, M. J., Robinson, E. B., Hyman, S. E., and Neale, B. M. (2019). Predicting polygenic risk of psychiatric disorders. *Biological psychiatry* **86,** 97–109.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51,** 584–591.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., Wang, Q., Bolla, M. K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics* **104,** 21–34.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., et al. (2016). Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature Neuroscience* **19,** 1523–1536.

Mistry, S., Harrison, J. R., Smith, D. J., Escott-Price, V., and Zammit, S. (2018). The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: A systematic review. *Journal of Affective Disorders* **234,** 148–155.

Ni, G., Moser, G., Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., et al. (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *The American Journal of Human Genetics* **102,** 1185–1194.

Nielsen, J. B., Thorolfsdottir, R. B., Fritsche, L. G., Zhou, W., Skov, M. W., Graham, S. E., Herron, T. J., McCarthy, S., Schmidt, E. M., Sveinbjornsson, G., et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature Genetics* **50,** 1234–1239.

Nikulin, V. V., Marzinzik, F., Wahl, M., Schneider, G.-H., Kupsch, A., Curio, G., and Klostermann, F. (2008). Anticipatory activity in the human thalamus is predictive of reaction times. *Neuroscience* **155,** 1275–1283.

Nivard, M. G., Gage, S. H., Hottenga, J. J., van Beijsterveldt, C. E., Abdellaoui, A., Bartels, M., Baselmans, B. M., Ligthart, L., Pourcain, B. S., Boomsma, D. I., et al. (2017). Genetic

overlap between schizophrenia and developmental psychopathology: longitudinal and multivariate polygenic risk prediction of common psychiatric traits during development. *Schizophrenia Bulletin* **43,** 1197–1207.

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics* **105,** 456–476.

Pasaniuc, B. and Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18,** 117–127.

Paul, D. and Aue, A. (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference* **150,** 1–29.

Pluta, D., Ombao, H., Chen, C., Xue, G., Moyzis, R., and Yu, Z. (2017). Adaptive mantel test for associationtesting in imaging genetics data. *arXiv preprint arXiv:1712.07270* .

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* **47,** 702–709.

Pouget, J. G., of the Psychiatric Genomics Consortium, S. W. G., Han, B., Wu, Y., Mignot, E., Ollila, H. M., Barker, J., Spain, S., Dand, N., Trembath, R., et al. (2019). Cross-disorder analysis of schizophrenia and 19 immune-mediated diseases identifies shared genetic risk. *Human molecular genetics* **28,** 3498–3513.

Power, R. A., Steinberg, S., Bjornsdottir, G., Rietveld, C. A., Abdellaoui, A., Nivard, M. M., Johannesson, M., Galesloot, T. E., Hottenga, J. J., Willemsen, G., et al. (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience* **18,** 953–955.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81,** 559–575.

Purcell, S. M., Wray, R., Stone, L., Visscher, M., O'Donovan, C., Sullivan, F., Sklar, P., Ruderfer, M., McQuillin, A., Morris, W., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752.

Quick, C., Fuchsberger, C., Taliun, D., Abecasis, G., Boehnke, M., and Kang, H. M. (2018). emerald: rapid linkage disequilibrium estimation with massive datasets. *Bioinformatics* **35,** 164–166.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science* **6,** 15–32.

Schaid, D., Chen, W., and Larson, N. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19,** 491–504.

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics* **101,** 737–751.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis* **55,** 331–339.

Socrates, A., Bond, T., Karhunen, V., Auvinen, J., Rietveld, C., Veijola, J., Jarvelin, M.-R., and O'Reilly, P. (2017). Polygenic risk scores applied to a single cohort reveal pleiotropy among hundreds of human phenotypes. *BioRxiv 203257* .

Somerville, L. H., Bookheimer, S. Y., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Dapretto, M., Elam, J. S., Gaffrey, M. S., Harms, M. P., Hodge, C., et al. (2018). The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *NeuroImage* **183,** 456–468.

Speed, D. and Balding, D. (2014). Multiblup: improved snp-based prediction for complex traits. *Genome Research* **24,** 1550–1557.

Speed, D. and Balding, D. (2019). Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature Genetics* **51,** 277–284.

Steinsaltz, D., Dahl, A., and Wachter, K. W. (2018). Statistical properties of simple random-effects models for genetic heritability. *Electronic Journal of Statistics* **12,** 321–358.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* **12,**

e1001779.

Sugrue, L. P. and Desikan, R. S. (2019). What are polygenic scores and why are they important? *JAMA* **321,** 1820–1821.

Sullivan, P. F. and Geschwind, D. H. (2019). Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. *Cell* **177,** 162–183.

Sun, R. and Lin, X. (2017). Set-based tests for genetic association using the generalized berk-jones statistic. *arXiv preprint arXiv:1710.02469* .

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20,** 467–484.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

Timpson, N. J., Greenwood, C. M., Soranzo, N., Lawson, D. J., and Richards, J. B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics* **19,** 110–125.

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19,** 581–590.

Tulino, A. M. and Verdú, S. (2004). Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory* **1,** 1–182.

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., Kandel, B. M., van Strien, N., Stone, J. R., Gee, J. C., et al. (2014). Large-scale evaluation of ants and freesurfer cortical thickness measurements. *Neuroimage* **99,** 166–179.

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics* **20,** 567–581.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* **97,** 576–592.

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A., Chen, G.-B., Lee, S. H., Wray, N. R., Goddard, M. E., and Yang, J. (2014). Statistical power to detect genetic (co) variance of complex traits using snp data in unrelated samples. *PLoS Genetics* **10,** e1004269.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics* **101,** 5–22.

Vreeker, A., Abramovic, L., Boks, M. P., Verkooijen, S., van Bergen, A. H., Ophoff, R. A., Kahn, R. S., and van Haren, N. E. (2017). The relationship between brain volumes and intelligence in bipolar disorder. *Journal of Affective Disorders* **223,** 59–64.

Wang, C., Pan, G., Tong, T., and Zhu, L. (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica* **25,** 993–1008.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78,** 589–611.

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., Polderman, T. J., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2018). A global view of pleiotropy and genetic architecture in complex traits. *bioRxiv 500090* .

Wei, Y., de Lange, S. C., Scholtens, L. H., Watanabe, K., Ardesch, D. J., Jansen, P. R., Savage, J. E., Li, L., Preuss, T. M., Rilling, J. K., Posthuma, D., and van den Heuvel, M. P. (2019). Genetic correlates of evolutionary adaptations in cognitive functional brain networks and their relationship to human cognitive functioning and disease. *bioRxiv* .

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. (2013). The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia* **9,** e111–e194.

Wheeler, H. E., Aquino-Michaels, K., Gamazon, E. R., Trubetskoy, V. V., Dolan, M. E., Huang, R. S., Cox, N. J., and Im, H. K. (2014). Poly-omic prediction of complex traits:

Omickriging. *Genetic Epidemiology* **38,** 402–415.

Wolff, M. and Vann, S. D. (2019). The cognitive thalamus as a gateway to mental representations. *Journal of Neuroscience* **39,** 3–14.

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J., and Visscher, P. M. (2018). Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173,** 1573–1580.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature Genetics* **42,** 565–569.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82.

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of snp-based heritability. *Nature Genetics* **49,** 1304–1310.

Yang, Q. and Cheng, G. (2018). Quadratic discriminant analysis under moderate dimension. *arXiv preprint arXiv:1808.10065* .

Yao, J., Zheng, S., and Bai, Z. (2015). *Sample covariance matrices and high-dimensional data analysis*, volume 2. Cambridge University Press Cambridge.

Zhao, B., Ibrahim, J. G., Li, Y., Li, T., Wang, Y., Shan, Y., Zhu, Z., Zhou, F., Zhang, J., Huang, C., et al. (2018). Heritability of regional brain volumes in large-scale neuroimaging and genetic studies. *Cerebral Cortex* **29,** 2904–2914.

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., and Zhu, H. (2019). Gwas of 19,629 individuals identifies novel genetic variants for regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *bioRxiv 586339* .

Zhao, B. and Zou, F. (2019). On prs for complex polygenic trait prediction. *bioRxiv 447797* .

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine*

Learning Research **7,** 2541–2563.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2018). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652* .

Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics* **9,** e1003264.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320.